

PATENT ABSTRACTS OF JAPAN

(11)Publication number : 2004-030344

(43)Date of publication of application : 29.01.2004

(51)Int.Cl. G06F 17/28

(21)Application number : 2002-186965 (71)Applicant : NIPPON TELEGR & TELEPH CORP <NTT>

(22)Date of filing : 26.06.2002 (72)Inventor : ADACHI TAKAYUKI
UCHINO HAJIME

(54) METHOD, DEVICE AND PROGRAM FOR MAKING TRANSLATION RULE

(57)Abstract:

PROBLEM TO BE SOLVED: To make a reliable rule for template translation.

SOLUTION: Variable candidates of parallel translation to each language and phrase correspondences between both languages are calculated by a plurality of methods (ST200), a combination to be appropriate correspondence in a portion of a sentence or the whole sentence is selected from the obtained phrase

se correspondences (ST300), certainty is calculated from the recognizing method and associating method f

the variable candidates and a word correspondence ratio of the whole sentence in order to determine whether a rule can be used for template translation (ST400), and a rule with high certainty is made s an appropriate template rule (ST500).

(19) 日本国特許庁(JP)

(12) 公開特許公報(A)

(11) 特許出願公開番号

特開2004-30344

(P2004-30344A)

(43) 公開日 平成16年1月29日(2004.1.29)

(51) Int. Cl.⁷
G06F 17/28F I
G O 6 F 17/28テーマコード (参考)
5 B O 9 1

審査請求 未請求 請求項の数 9 O L (全 35 頁)

(21) 出願番号	特願2002-186965 (P2002-186965)	(71) 出願人	000004226
(22) 出願日	平成14年6月26日 (2002. 6. 26)		日本電信電話株式会社
			東京都千代田区大手町二丁目 3 番 1 号
		(74) 代理人	100058479
			弁理士 鈴江 武彦
		(74) 代理人	100084618
			弁理士 村松 貞男
		(74) 代理人	100087963
			弁理士 石川 義雄
		(72) 発明者	足立 貴行
			東京都千代田区大手町二丁目 3 番 1 号 日
			本電信電話株式会社内
		(72) 発明者	内野 一
			東京都千代田区大手町二丁目 3 番 1 号 日
			本電信電話株式会社内
		F ターム (参考)	5B091 CA21 CC03

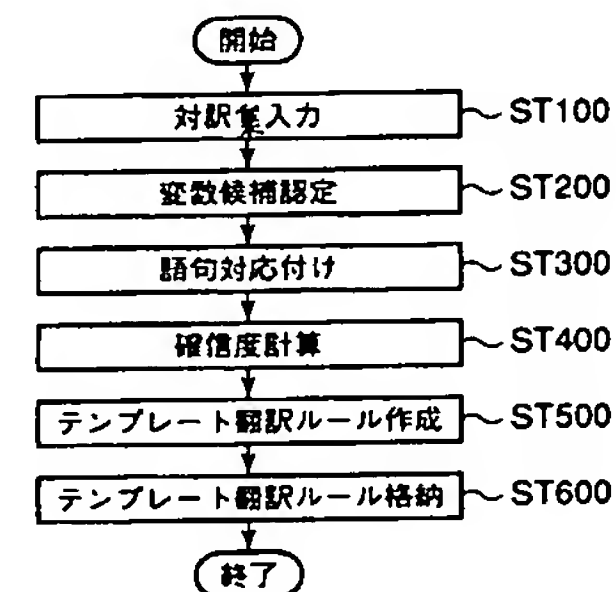
(54) 【発明の名称】 翻訳ルール作成方法、装置およびプログラム

(57) 【要約】

【課題】 信頼性の高いテンプレート翻訳用ルールを作成する。

【解決手段】 対訳对各言語の変数候補および両言語間での語句対応を複数の手法によって求め (S T 2 0 0)、得られた語句対応から文の一部および文全体で適切な対応となる組み合わせを選び (S T 3 0 0)、テンプレート翻訳に利用可能なルールか判定するために、変数候補の認定手法や対応付け手法や文全体の単語対応割合から確信度を計算して (S T 4 0 0)、確信度の高いものを適切なテンプレート翻訳ルールとして作成する (S T 5 0 0)。

【選択図】 図 2



【特許請求の範囲】

【請求項 1】

第 1 自然言語文とその第 1 自然言語文の訳である第 2 自然言語文との対である対訳対をもとにして翻訳ルールを作成する翻訳ルール作成方法において、
前記第 1 および第 2 自然言語文に対して単語とその単語の属性情報とが関連づけられている属性辞書を参照して、前記第 1 自然言語文を構成している単語とその単語の品詞を含む単語の属性情報を抽出するステップと、
前記属性辞書を参照して、前記第 2 の自然言語文を構成している単語とその単語の品詞を含む属性情報を抽出するステップと、
前記抽出された属性情報にもとづいて、所定の属性を含む単語を置換可能候補として決定するステップと、
前記第 1 自然言語文と前記第 2 自然言語文との間で、それら各文を構成する単語の単語表記対、文字対、および、音韻対のうち、少なくともいずれかが格納された対辞書を参照して、前記第 1 自然言語文の単語に対応する前記第 2 自然言語文の単語を抽出し、前記第 2 自然言語文を構成している単語のうち前記抽出された単語に一致する単語を対応付け候補として決定するステップと、
前記対辞書に格納されている前記単語表記対、前記文字対、および、前記音韻対のうちのいずれにもとづいて対応付けがなされるかに対応して数値が付与されている重み辞書を参照して、前記対応付け候補に対応度値を付与するステップと、
互いに単語が重複して選択されないように前記対応付け候補の組み合わせごとに対応度値にもとづいて、文単位の対応度を計算するステップと、
前記文単位の対応度が最大となる対応付けの組み合わせを抽出するステップと、
各自然言語文の対応付けられた単語数またはその単語の属性にもとづいて確信度値を計算するステップと、
ある閾値以上の確信度をもつ対応づけの両言語文を構成する単語、その単語の属性、および、置換可能候補からなる翻訳ルール情報を記憶手段に記憶するステップと、
を具備することを特徴とする翻訳ルール作成方法。

10

20

30

40

50

【請求項 2】

前記確信度値を計算するステップでは、前記第 1 および第 2 自然言語文のそれぞれに含まれている前記対応付け候補である単語のうち内容語である単語数と、前記第 1 および第 2 自然言語文に含まれている内容語である単語数との比にもとづいて確信度値を計算することを特徴とする請求項 1 に記載の翻訳ルール作成方法。

【請求項 3】

前記確信度値を計算するステップでは、前記第 1 および第 2 自然言語文のそれぞれに含まれている置換可能候補である単語の属性情報に対応して付加されている重み値の和を、前記第 1 および第 2 自然言語文のそれぞれに含まれている置換可能候補数で除した値にもとづいて確信度値を計算することを特徴とする請求項 1 に記載の翻訳ルール作成方法。

【請求項 4】

第 1 自然言語文とその第 1 自然言語文の訳である第 2 自然言語文との対である対訳対をもとにして翻訳ルールを作成する翻訳ルール作成装置において、
前記第 1 および第 2 自然言語文に対して単語とその単語の属性情報とが関連づけられている属性辞書を参照して、前記第 1 自然言語文を構成している単語とその単語の品詞を含む単語の属性情報を抽出する第 1 抽出手段と、
前記属性辞書を参照して、前記第 2 の自然言語文を構成している単語とその単語の品詞を含む属性情報を抽出する第 2 抽出手段と、
前記抽出された属性情報にもとづいて、所定の属性を含む単語を置換可能候補として決定する置換可能候補決定手段と、
前記第 1 自然言語文と前記第 2 自然言語文との間で、それら各文を構成する単語の単語表記対、文字対、および、音韻対のうち、少なくともいずれかが格納された対辞書を参照して、前記第 1 自然言語文の単語に対応する前記第 2 自然言語文の単語を抽出し、前記第 2

自然言語文を構成している単語のうち前記抽出された単語に一致する単語を対応付け候補として決定する対応付け手段と、
前記対辞書に格納されている前記単語表記対、前記文字対、および、前記音韻対のうちのいずれにもとづいて対応付けがなされるかに対応して数値が付与されている重み辞書を参照して、前記対応付け候補に対応度値を付与する対応度手段と、
互いに単語が重複して選択されないように前記対応付け候補の組み合わせごとに対応度値にもとづいて、文単位の対応度を計算する対応度計算手段と、
前記文単位の対応度が最大となる対応付けの組み合わせを抽出する第3抽出手段と、
各自然言語文の対応付けられた単語数またはその単語の属性にもとづいて確信度値を計算する確信度計算手段と、
ある閾値以上の確信度をもつ対応づけの両言語文を構成する単語、その単語の属性、および、置換可能候補からなる翻訳ルール情報を記憶する記憶手段と、
を具備することを特徴とする請求項に記載の翻訳ルール作成装置。

10

【請求項5】

前記確信度計算手段は、前記第1および第2自然言語文のそれぞれに含まれている前記対応付け候補である単語のうち内容語である単語数と、前記第1および第2自然言語文に含まれている内容語である単語数との比にもとづいて確信度値を計算することを特徴とする請求項4に記載の翻訳ルール作成装置。

【請求項6】

前記確信度値を計算するステップでは、前記第1および第2自然言語文のそれぞれに含まれている置換可能候補である単語の属性情報に対応して付加されている重み値の和を、前記第1および第2自然言語文のそれぞれに含まれている置換可能候補数で除した値にもとづいて確信度値を計算することを特徴とする請求項4に記載の翻訳ルール作成装置。

20

【請求項7】

コンピュータに、第1自然言語文とその第1自然言語文の訳である第2自然言語文との対である対訳対をもとにして翻訳ルールを作成するためのステップを実行させる翻訳ルール作成プログラムにおいて、

前記第1および第2自然言語文に対して単語とその単語の属性情報とが関連づけられている属性辞書を参照して、前記第1自然言語文を構成している単語とその単語の品詞を含む単語の属性情報を抽出するステップと、

30

前記属性辞書を参照して、前記第2の自然言語文を構成している単語とその単語の品詞を含む属性情報を抽出するステップと、

前記抽出された属性情報にもとづいて、所定の属性を含む単語を置換可能候補として決定するステップと、

前記第1自然言語文と前記第2自然言語文との間で、それら各文を構成する単語の単語表記対、文字対、および、音韻対のうち、少なくともいずれかが格納された対辞書を参照して、前記第1自然言語文の単語に対応する前記第2自然言語文の単語を抽出し、前記第2自然言語文を構成している単語のうち前記抽出された単語に一致する単語を対応付け候補として決定するステップと、

前記対辞書に格納されている前記単語表記対、前記文字対、および、前記音韻対のうちのいずれにもとづいて対応付けがなされるかに対応して数値が付与されている重み辞書を参照して、前記対応付け候補に対応度値を付与するステップと、

40

互いに単語が重複して選択されないように前記対応付け候補の組み合わせごとに対応度値にもとづいて、文単位の対応度を計算するステップと、

前記文単位の対応度が最大となる対応付けの組み合わせを抽出するステップと、

各自然言語文の対応付けられた単語数またはその単語の属性にもとづいて確信度値を計算するステップと、

ある閾値以上の確信度をもつ対応づけの両言語文を構成する単語、その単語の属性、および、置換可能候補からなる翻訳ルール情報を記憶手段に記憶するステップと、

を実行させることを特徴とする翻訳ルール作成プログラム。

50

【請求項 8】

前記確信度値を計算するステップでは、前記第 1 および第 2 自然言語文のそれぞれに含まれている前記対応付け候補である単語のうち内容語である単語数と、前記第 1 および第 2 自然言語文に含まれている内容語である単語数との比にもとづいて確信度値を計算することを特徴とする請求項 7 に記載の翻訳ルール作成プログラム。

【請求項 9】

前記確信度値を計算するステップでは、前記第 1 および第 2 自然言語文のそれぞれに含まれている置換可能候補である単語の属性情報に対応して付加されている重み値の和を、前記第 1 および第 2 自然言語文のそれぞれに含まれている置換可能候補数で除した値にもとづいて確信度値を計算することを特徴とする請求項 7 に記載の翻訳ルール作成プログラム

10

【発明の詳細な説明】

【0001】

【発明の属する技術分野】

本発明のテンプレート翻訳ルール作成方法、装置およびプログラムは、第 1 自然言語文とその訳の第 2 自然言語文との対である対訳対を用いて、テンプレート翻訳用ルールを作成する方法、装置およびプログラムに関するものである。

【0002】

【従来の技術】

従来技術 1 として、対訳対に対して対訳辞書と句解析処理結果から語句対応付けを行い、対応箇所を変数化して翻訳テンプレートを作成する特開平 5-151260 の「翻訳テンプレート学習方法および翻訳テンプレート学習システム」がある。

20

【0003】

また、従来技術 2 として、対訳対の単語列を分類記号に変換したものをパターンとして事前に作成し、翻訳の際には入力文を同様にパターン化し、事前作成されたパターンとの照合やパターン内の変数への代入により翻訳を行う特開平 5-233693 の「機械翻訳方式」がある。

【0004】

【発明が解決しようとする課題】

上述の従来技術 1 では、対応した単語やそれを包含する句をテンプレートの変数としており、対応付けには対訳辞書を用いているので、対訳辞書にない場合は対応が得られないという問題がある。

30

【0005】

また、上述の従来技術 2 では、変数を数詞などの定型的な単語列の対応や局所辞書にある単語列の対応としているが、定型的でなく、局所辞書にない単語列は、変数とすることができないという問題がある。

【0006】

さらに、従来手法 1、2 では、自動作成されたテンプレートについて、翻訳するのに適しているか否か判定することは考慮されていないので、生成されたテンプレートの信頼性が低い場合があるという問題がある。

40

【0007】

そこで本発明は、上述した従来の問題点に鑑み、信頼性の高いテンプレート翻訳用ルールを作成する翻訳ルール作成方法、装置およびプログラムを提供することを目的とする。

【0008】

【課題を解決するための手段】

本発明の翻訳ルール作成方法は、
第 1 自然言語文とその第 1 自然言語文の訳である第 2 自然言語文との対である対訳対をもとにして翻訳ルールを作成する翻訳ルール作成方法において、
第 1 および第 2 自然言語文に対して単語とその単語の属性情報とが関連づけられている属性辞書を参照して、前記第 1 自然言語文を構成している単語とその単語の品詞を含む単語

50

の属性情報を抽出するステップと、

前記属性辞書を参照して、前記第2の自然言語文を構成している単語とその単語の品詞を含む属性情報を抽出するステップと、

前記抽出された属性情報にもとづいて、所定の属性を含む単語を置換可能候補として決定するステップと、

第1自然言語文と第2自然言語文との間で、それら各文を構成する単語の単語表記対、文字対、および、音韻対のうち、少なくともいずれかが格納された対辞書を参照して、第1自然言語文の単語に対応する第2自然言語文の単語を抽出し、第2自然言語文を構成している単語のうち前記抽出された単語に一致する単語を対応付け候補として決定するステップと、

10

前記対辞書に格納されている前記単語表記対、前記文字対、および、前記音韻対のうちのいずれにもとづいて対応付けがなされるかに対応して数値が付与されている重み辞書を参照して、前記対応付け候補に対応度値を付与するステップと、

互いに単語が重複して選択されないように前記対応付け候補の組み合わせごとに対応度値にもとづいて、文単位の対応度を計算するステップと、

前記文単位の対応度が最大となる対応付けの組み合わせを抽出するステップと、

各自然言語文の対応付けられた単語数またはその単語の属性にもとづいて確信度値を計算するステップと、

ある閾値以上の確信度をもつ対応づけの両言語文を構成する単語、その単語の属性、および、置換可能候補からなる翻訳ルール情報を記憶手段に記憶するステップとを備えている

20

。

【0009】

また、本発明の翻訳ルール作成装置は、

第1自然言語文とその第1自然言語文の訳である第2自然言語文との対である対訳対をもとにして翻訳ルールを作成する翻訳ルール作成装置において、

第1および第2自然言語文に対して単語とその単語の属性情報とが関連づけられている属性辞書を参照して、前記第1自然言語文を構成している単語とその単語の品詞を含む単語の属性情報を抽出する第1抽出手段と、

前記属性辞書を参照して、前記第2の自然言語文を構成している単語とその単語の品詞を含む属性情報を抽出する第2抽出手段と、

30

前記抽出された属性情報にもとづいて、所定の属性を含む単語を置換可能候補として決定する置換可能候補決定手段と、

第1自然言語文と第2自然言語文との間で、それら各文を構成する単語の単語表記対、文字対、および、音韻対のうち、少なくともいずれかが格納された対辞書を参照して、第1自然言語文の単語に対応する第2自然言語文の単語を抽出し、第2自然言語文を構成している単語のうち前記抽出された単語に一致する単語を対応付け候補として決定する対応付け手段と、

前記対辞書に格納されている前記単語表記対、前記文字対、および、前記音韻対のうちのいずれにもとづいて対応付けがなされるかに対応して数値が付与されている重み辞書を参照して、前記対応付け候補に対応度値を付与する対応度手段と、

40

互いに単語が重複して選択されないように前記対応付け候補の組み合わせごとに対応度値にもとづいて、文単位の対応度を計算する対応度計算手段と、

前記文単位の対応度が最大となる対応付けの組み合わせを抽出する第3抽出手段と、

各自然言語文の対応付けられた単語数またはその単語の属性にもとづいて確信度値を計算する確信度計算手段と、

ある閾値以上の確信度をもつ対応づけの両言語文を構成する単語、その単語の属性、および、置換可能候補からなる翻訳ルール情報を記憶する記憶手段とを備えている。

【0010】

さらに、翻訳ルール作成プログラムは、

コンピュータに、第1自然言語文とその第1自然言語文の訳である第2自然言語文との対

50

である対訳対をもとにして翻訳ルールを作成するためのステップを実行させる翻訳ルール作成プログラムにおいて、

第1および第2自然言語文に対して単語とその単語の属性情報とが関連づけられている属性辞書を参照して、前記第1自然言語文を構成している単語とその単語の品詞を含む単語の属性情報を抽出するステップと、

前記属性辞書を参照して、前記第2の自然言語文を構成している単語とその単語の品詞を含む属性情報を抽出するステップと、

前記抽出された属性情報にもとづいて、所定の属性を含む単語を置換可能候補として決定するステップと、

第1自然言語文と第2自然言語文との間で、それら各文を構成する単語の単語表記対、文字対、および、音韻対のうち、少なくともいずれかが格納された対辞書を参照して、第1自然言語文の単語に対応する第2自然言語文の単語を抽出し、第2自然言語文を構成している単語のうち前記抽出された単語に一致する単語を対応付け候補として決定するステップと、 10

前記対辞書に格納されている前記単語表記対、前記文字対、および、前記音韻対のうちのいずれにもとづいて対応付けがなされるかに対応して数値が付与されている重み辞書を参照して、前記対応付け候補に対応度値を付与するステップと、

互いに単語が重複して選択されないように前記対応付け候補の組み合わせごとに対応度値にもとづいて、文単位の対応度を計算するステップと、

前記文単位の対応度が最大となる対応付けの組み合わせを抽出するステップと、 20

各自然言語文の対応付けられた単語数またはその単語の属性にもとづいて確信度値を計算するステップと、

ある閾値以上の確信度をもつ対応づけの両言語文を構成する単語、その単語の属性、および、置換可能候補からなる翻訳ルール情報を記憶手段に記憶するステップとを実行させる。

【0011】

以上の構成によれば、対訳対を利用したテンプレート翻訳ルールの作成において、対訳対の各言語の変数候補および両言語間での語句対応を複数の手法によって求め、得られた語句対応から文の一部および文全体で適切な対応となる組み合わせを選び、テンプレート翻訳に利用可能なルールか判定するために、変数候補の認定手法や対応付け手法や文全体の単語対応割合から確信度を計算して、確信度の高いものを適切なテンプレート翻訳ルールとして作成することができる。 30

【0012】

【発明の実施の形態】

以下、図面を参照して本発明の実施形態の翻訳ルール作成方法、装置およびプログラムについて説明する。

【0013】

図1は、本発明の実施形態に係る翻訳ルール作成装置1000の機能ブロック図である。本実施形態に係る翻訳ルール作成装置1000は、インタフェース部1100、テンプレート翻訳ルール作成部1200、テンプレート翻訳ルール作成制御部1300、テンプレート翻訳ルールデータベース1400、およびメモリ1500を備えている。 40

【0014】

インタフェース部1100は、第1の自然言語文およびその文の第2の自然言語文への対訳文を入力する。入力された対訳文は、テンプレート翻訳ルール作成制御部1300を介してメモリ1500に出力される。テンプレート翻訳ルール作成制御部1300は、全体の制御手段である。メモリ1500は、対訳文を格納する。さらに、メモリ1500は、各種のデータベースを格納している。テンプレート翻訳ルール作成部1200は、メモリ1500から対訳文を参照して、この対訳文にもとづいて翻訳のためのテンプレートを作成する。

テンプレート翻訳ルール作成制御部1300は、各部、メモリ、データベース間でのデー 50

タの入出力を制御するとともに、テンプレート翻訳ルールを作成するための動作を制御する。

【0015】

テンプレート翻訳ルールデータベース1400は、テンプレート翻訳ルール作成部1200によって作成されたテンプレート翻訳ルールを格納する。

【0016】

図2は、本発明の実施形態に係る翻訳ルール作成装置1000で使用する翻訳ルール作成方法における処理の流れを示すフローチャートである。

【0017】

第1の自然言語文およびその文の第2の自然言語文への対訳文が格納されている対訳集が、テンプレート翻訳ルール作成制御部1300の制御の下にインタフェース部1100からメモリ1500に格納される。(ST100)。

【0018】

テンプレート翻訳ルール作成制御部1300の制御の下に各対訳文がメモリ1500からテンプレート翻訳ルール作成部1200に入力される。

【0019】

ステップST100において入力された各対訳対に対して、固有名詞、数詞等の属性を有する単語が抽出され、これら抽出された単語が変数候補と決定される(ST200)。以下ステップST200からステップST500までは、テンプレート翻訳ルール作成部1200が実行するステップである。

【0020】

ここで変数候補とは、翻訳時にその翻訳内容にしたがって置換可能である候補を意味する。

【0021】

なお、変数候補の対象を固有名詞や数詞としているが、それ以外の語句でも同様な処理であればこれらに限定されない。

【0022】

ステップST200で決定された変数候補である単語およびそのほかの各対訳対の単語を両言語間で対応付けが実行される(ST300)。ステップST300の語句対応付けでは、認定した変数候補およびすべての語句対応付けを行って対応度が最大のものを求める。

【0023】

ステップST400の確信度計算では変数候補認定や対応付け手法や単語対応割合から確信度を求める。

【0024】

ステップST500のテンプレート翻訳ルール作成では、対訳対から変数候補や対応のデータを利用してテンプレート翻訳のルールを作成する。

ステップST600のテンプレート翻訳ルール格納では、テンプレート翻訳ルール作成制御部1300の制御下で、作成されたテンプレート翻訳ルールがテンプレート翻訳ルールデータベース1400に格納される。

【0025】

図3は、図2のステップST200の変数候補認定の処理の流れを示すフローチャートである。以下のフローチャートは、テンプレート翻訳ルール作成制御部1300の指示にもとづいて実行される。

ステップST210で形態素解析が利用できるかどうか判定する。形態素解析が利用できる場合はステップST220の形態素解析で入力文の形態素解析を行い固有名詞や数詞を認定する。ステップST210で形態素解析が利用できない場合はステップST230の単語切りで文を単語分割する。

【0026】

ステップST240の固有名詞辞書検索で入力文の語句に対して固有名詞辞書を検索して

10

20

30

40

50

固有名詞を認定し、変数候補とする。

ステップ S T 2 5 0 の数詞解析でアラビア数字列ではないが数字として扱える文字列をアラビア数字列に変換し、数詞と認定し、変数候補とする。

ステップ S T 2 6 0 で単語表記から固有名詞とわかるか否かを調べる。固有名詞とわかる場合は、ステップ S T 2 7 0 の固有名詞表記抽出で表記から固有名詞を抽出し、認定する。また、固有名詞を変数候補とする。

【 0 0 2 7 】

ステップ S T 2 2 0 の形態素解析が終わった後、またはステップ S T 2 6 0 で表記から固有名詞とわからなかった後、またはステップ S T 2 7 0 の固有名詞表記抽出が終わった後、ステップ S T 2 8 0 のパターンマッチで、パターンとのマッチングを行って変数候補を認定する。 10

【 0 0 2 8 】

図 4 は、図 3 のステップ S T 3 0 0 の処理の流れを示すフローチャートである。以下のフローチャートは、テンプレート翻訳ルール作成制御部 1 3 0 0 の指示にもとづいて実行される。

ステップ S T 3 1 0 の対訳辞書対応では、対訳辞書を用いて対訳間の対応を求める。

ステップ S T 3 2 0 の文字対応では、両言語で同じ文字や異体字の対応を求める。

ステップ S T 3 3 0 の類似音韻対応では、両言語間の類似する音韻によって対応を求める。

ステップ S T 3 4 0 のパターン対応では、ステップ S T 2 8 0 のパターンマッチで作成されたパターンの標準形が両言語間で一致するものを求める。 20

【 0 0 2 9 】

ステップ S T 3 5 0 の最適対応選択では、得られた対応候補の情報を利用して最適な対応を選択する。

ステップ S T 3 5 1 の最適部分対応では、文中のある区間に限定して、区間中の全ての対応から重複のない最適な対応を求める。

ステップ S T 3 5 2 の最適文対応では、ステップ S T 3 5 1 で得られた最適な対応から文全体で最適な対応となる語句対応の組み合わせを求める。

【 0 0 3 0 】

上述したフロー図の各ステップに示された指示は、ソフトウェアであるプログラムにもとづいて実行される。プログラムは、本実施形態の翻訳ルール作成装置であるプログラム可能な装置上でロードされる。この翻訳ルール作成装置上で実行される指示は、フロー図の各ブロックで特徴づけられている機能を実行する手段を提供する。プログラムは、メモリ 1 5 0 0、これに類する記録媒体である C D - R O M 等に格納、または、通信回線によって C P U 等の演算手段を介してインストールされて本発明の本実施形態の方法が実行される。 30

【 0 0 3 1 】

【実施例】

以下、図面を参照して本発明の実施例の翻訳ルール作成方法、装置およびプログラムについて説明する。 40

【 0 0 3 2 】

(実施例 1)

以下、対訳間の単語対応割合により、テンプレート翻訳に適したルールを選別できることを示す。本実施例では第 1 自然言語を日本語、第 2 自然言語を英語とする。

図 5 は、本発明の実施例 1 において入力される対訳対 I D が 1 の対訳集である。

インタフェース部 1 1 0 0 は、図 5 に示されている対訳集を入力する。入力された対訳集は、メモリ 1 5 0 0 に格納される (図 2 の S T 1 0 0)。テンプレート翻訳ルール作成部 1 2 0 0 では、メモリ 1 5 0 0 に格納されている図 5 の対訳集から各対訳対を対訳対 I D 順に上述した図 2 の S T 2 0 0 から S T 5 0 0 に示された処理を行う。以下、はじめに対訳対の言語が日本語の文の場合を説明し、つぎにもう 1 つの対訳対の言語が英語の文の場合 50

合を説明する。

【0033】

まず、図5の対訳集から対訳対ID=1の対訳対を読み込む。読み込まれた対訳対に対して、形態素解析が利用できるか否かが判定される(図3のST210)。

【0034】

日本語文について、形態素解析手段が用意されていたとすると、形態素解析が利用できるとして、形態素解析が実行される(ST210、ST220)。形態素解析手段は、言語ごとに予め用意されていたり、予め用意されていなかったりする。ここでは、日本語の文について、予め形態素解析手段が用意されていたということである。

【0035】

また、形態素解析手段とは、テキスト情報からその情報を構成する単語、その単語の品詞、その単語の読み、およびほかの単語との係り受け情報を解析するものである。この形態素解析手段は公知の手段であり、たとえば、「宮崎正弘ら：日本文音声出力のための言語処理方式、情報処理学会論文誌、Vol. 27、No. 11、pp1053-1061、1986」にその手段の記載がある。

【0036】

図6は、本発明の実施例1において日本語文に対して形態素解析(ST220)を実行した場合の解析結果である。この解析結果は、メモリ1500に格納される。

本実施例での形態素解析では、図5の対訳集の対訳対ID=1の言語欄に記載されている言語である“日”の文を形態素解析する。図6に示されている対訳対ID=1の言語である“日”の単語欄に各単語表記を格納する。各単語表記に対応して単語ID(文節-単語)欄に“文節番号-文節内の単語番号”を格納する。

【0037】

さらに、解析情報欄に品詞と読みを格納する。たとえば、単語ID=1-1の解析情報欄に品詞である“固有名詞”を格納し、読みである“nippon”を格納する。

【0038】

変数候補欄には、上記の結果をもとに変数として扱う単語に、その単語の品詞と対応している符号が格納される。本実施例では、単語ID=1-1の変数候補欄に固有名詞を示す符号“PN”が格納される。さらに、単語ID=1-1の変数候補欄に形態素解析による結果であることを示す符号“形”が格納される。

【0039】

また、単語ID=2-2、4-1、および、4-3の解析情報欄に“数詞”が格納されているので、これら単語IDの変数候補欄に数詞を示す符号“NUM”が格納される。また、単語ID=1-1の変数候補欄に形態素解析による結果であることを示す符号“形”が格納される。

【0040】

図7は、図6の作成後、ステップST280で参照されるパターン集である。このパターン集は、メモリ1500に格納されている。

ステップST220の形態素解析が完了した後、図7に示されるパターン集が参照されて、対訳対ID=1の日本語文のうち図7に示されるパターンに合う日本語文中の単語があるかが調べられる(ST280)。

【0041】

pID欄は、パターンごとの識別番号を格納する。言語欄は、pIDで指定されるパターンがどの言語の文に適用されるかを示す。図7によれば、本実施例では、パターン集の中のパターンID(pID)がp1およびp2である場合が日本語文に対して適用されることがわかる。

【0042】

標準形欄は、後のパターン対応で利用するもので、パターン中の両言語の表記を共通化するルールを格納する。なお、パターンの標準形は表記と“\$番号”の並びであり、“\$番号”はP単語ID欄と同じ番号で一致した単語(品詞が数詞の場合は半角のアラビア数字

10

20

30

40

50

に変換されている)に置換される。

P単語ID欄は、変数候補の位置を指定する番号が格納されている。単語欄は、一致するパターンの単語が格納されている。単語欄が空白である場合は、単語は指定されていないことを示している。

【0043】

品詞欄は、パターンの品詞が格納されている。変数候補欄は、図6に示されるような解析結果の変数候補欄の記号が格納される。この変数候補欄に格納されている品詞を示す記号に対応する単語が変数候補になる。

【0044】

たとえば、ステップST280では、PID = p1に関しては、図6の解析結果の対訳対ID = 1において、単語ID = 2 - 2、4 - 1、4 - 3と一致する。また、パターン集の変数候補欄が“NUM”となっているので、図6解析結果の対訳対ID = 1の各単語ID = 2 - 2、4 - 1、4 - 3の変数候補欄に“NUM”を格納し、変数候補認定がパターンによるので変数候補欄に“パ”を格納する(図9)。

10

【0045】

図8は、図7のパターン集にもとづいて抽出されたパターン情報である。このパターン情報は、メモリ1500に格納される。

図7のパターン集を参照して、図6に示される解析結果から、パターン集のパターンに対応する先頭の単語ID、パターン集のパターンに対応する末尾の単語ID、その単語の標準形をその単語に対応させた具体的な形式に変換したもの、および、そのパターン集の中の変数候補欄に格納されている記号に対応している、解析結果の変数候補欄に含まれる単語IDの4つがパターン情報として抽出される(ST280)。

20

【0046】

ただし、パターンが1単語からなる場合は、パターン集のパターンに対応する先頭の単語IDとパターン集のパターンに対応する末尾の単語IDとは同一の単語IDとなるので、この場合は、パターン集の変数候補に対応する末尾の単語IDを省略してパターン情報として抽出される。

【0047】

たとえば、図8のパターン情報のIDが1の場合では、パターン集のパターンに対応する先頭の単語IDと末尾の単語IDは、それぞれ2 - 2および2 - 2である。また、その単語の標準形をその単語に対応させた具体的な形式に変換したものは、4である。さらに、パターン集の中の変数候補欄に格納されている記号に対応している、解析結果の変数候補欄に含まれる単語IDは、2 - 2である。

30

【0048】

この場合、パターンの標準形が1単語であるので、パターン集の変数候補に対応する先頭の単語IDが省略されて、情報欄には(2 - 2 : 4 : 2 - 2)が格納される。

【0049】

ほかに、たとえば、文中に「/6/時/30/分/」(/は単語区切りを示すとする)があり、図7のパターン集PID = p2と一致する場合は、パターンのP単語ID = 1は“6”、P単語ID = 2は“時”、P単語ID = 3は“30”、P単語ID = 4は“分”と一致することになる。その結果、標準形は\$1を“6”、\$3を“30”に置換した“6時30分”となる。

40

【0050】

また、パターン中の変数候補と一致した単語IDについて、図7に示されているPID = p2のように変数候補が複数含まれているパターンと一致した場合は、その対応するパターン情報の欄は、上記の「/6/時/30/分/」の例では、たとえば(1 - 1, 1 - 4 : 6時30分 : 1 - 1, 1 - 3)となる。この例では、「6」に単語ID 1 - 1、「時」に単語ID 1 - 2、「30」に単語ID 1 - 3、「分」に単語ID 1 - 4が付与されている仮定している。

【0051】

50

この例の（１－１，１－４：６時３０分：１－１，１－３）の「１－１，１－４」は、パターンの範囲が単語ＩＤ１－１から単語ＩＤ１－４であることを示している。「６時３０分」は標準形である。「１－１，１－３」は変数候補が単語ＩＤ１－１と単語ＩＤ１－３であることを示している。すなわち、「時」と「分」は固定されている。

【００５２】

また、たとえば、文中に「／６／時／３０／分／２０／秒／」等がある場合は、変数候補が「６」、「３０」、「２０」の位置になるので、この場合のパターン情報の欄中の変数候補を記載する欄には、単語ＩＤ番号が３つ格納されることになる。一般に、変数候補が複数になる場合は、それに対応した数の単語ＩＤ番号がパターン情報の変数候補を記載する欄に格納される。

10

【００５３】

したがって、ステップＳＴ２８０では図８のパターン情報の対訳対ＩＤ＝１の欄には、図７のパターン集のｐＩＤ＝ｐ１と一致した図６の解析結果の対訳対ＩＤ＝１の該当箇所を示す単語ＩＤ＝２－２、４－１、４－３と、パターンの標準形は“４”、“６”、“４”と、パターン中の変数候補となる箇所の単語ＩＤ＝２－２、４－１、４－３をそれぞれ格納する。

【００５４】

一方、英語文については、形態素解析が用意されていないとする。この場合、形態素解析が利用できないと判定される（ＳＴ２１０）。

20

【００５５】

図１０は、本発明の実施例１において英語文に対して単語ＩＤを付してその単語の読み付した解析結果である。この解析結果は、メモリ１５００に格納される。

【００５６】

ステップＳＴ２１０で形態素解析が利用できないと判定されると、対訳集にある英語文に関して単語切りが実行される（ＳＴ２３０）。すなわち、英語文に対して、コンマ、ピリオド、コロンのように単語間で空白がない場合は空白を挿入した後、空白を単語境界として単語間を分離して扱う。そして、単語欄に単語を、単語ＩＤ欄に“１－単語番号”文頭の単語から順に付与し、それらの単語ＩＤを単語ＩＤ欄に格納する（ＳＴ２３０）。これらコンマ等は、それらをパターンとして認識するパターン検出手段によって抽出される。抽出されたコンマ等は、単語とコンマ等との間に空白が挿入される。

30

【００５７】

単語切りは、言語にもとづいて処理が異なる。英語、ドイツ語、フランス語等のように、コンマ等の記号を除けば、単語間に空白があるので、その空白で単語を区切れればよい。一方、中国語、日本語の場合は、各文字で区切る。

【００５８】

また、英語では単語の読みは単語表記の綴りから推定できるので、この例では表記をそのまま解析情報欄に格納する（ＳＴ２３０）。

【００５９】

図１１は、英語単語とその品詞（前置詞、冠詞等）を対応づけてある付属語辞書である。この付属語辞書は、メモリ１５００に格納されている。図１２は、読みの付与および図１１の付属語辞書を参照して前置詞および冠詞を解析情報に付した解析結果である。この解析結果は、メモリ１５００に格納される。

40

図１１の付属語辞書は、実質的な意味内容に乏しい、文法機能を示す非自立語と品詞が対応付けてある辞書である。品詞としては、冠詞、前置詞、接続詞、関係詞、限定詞、および、助動詞などである。

この付属語辞書を参照して、図１０に示された解析結果の単語から、上記の冠詞、前置詞、接続詞、関係詞、限定詞、および、助動詞などの品詞を属性として有する単語を抽出して、図１２に示されるように解析情報欄にその品詞名を格納する（ＳＴ２３０）。

【００６０】

図１３は、英語の固有名詞をリスト化してある固有名詞辞書である。この固有名詞辞書は

50

、メモリ 1500 に格納されている。図 14 は、図 13 の固有名詞辞書を参照して固有名詞を解析情報に付してさらに変数候補を抽出した解析結果である。この解析結果は、メモリ 1500 に格納される。

図 13 の固有名詞辞書は、英語の固有名詞が列挙してある。この固有名詞辞書を参照して、英語文の任意の単語列について辞書に載っているか否かを調べそのうちの最も長い単語から選択される。(ST240)。

【0061】

その結果、図 12 に示された解析結果中の各単語が固有名詞であるか否かが判定される。固有名詞であると判定された単語が、図 12 に示された解析結果の単語から抽出されて、図 14 に示されるように抽出した単語に対応した解析情報欄に「固有名詞」を格納する (ST240)。

さらに、この「固有名詞」を付された単語を変数候補として、変数候補欄に固有名詞を示す「PN」を格納する (ST240)。さらにまた、この「固有名詞」を付された単語は、固有名詞辞書を参照して認定されたことを示すために変数候補欄に「辞」を格納する (ST240)。本実施例の図 12 の場合は、単語 ID = 1-1 が固有名詞であるとして抽出される。

【0062】

図 15 は、数字辞書およびパターン辞書を参照して数詞を解析情報に付してさらに変数候補を抽出した解析結果である。この解析結果は、メモリ 1500 に格納される。

図 14 に示される解析結果から、まだ変数候補となっていない単語に対して、数字を示す単語をパターンとして抽出する (ST250)。たとえば、「thousand」、「million」等が検索されれば、その単語およびその単語の直前に数詞があるはずであるとして「thousand」、「million」等を抽出する。また、1、2、3 等のアラビア数字、one、two、three 等の英数字、first、second、third 等の英序数詞等も数字を示すパターンとして抽出される。このパターン抽出は、言語ごとに用意されているパターン辞書を参照して実行される。このパターン辞書はメモリ 1500 に格納されている。

【0063】

このように抽出された単語は、数詞であることを示すために対応する解析情報欄に「数詞」が格納される (ST250)。さらに、この「数詞」を付された単語を変数候補として、変数候補欄に数詞を示す「NUM」が格納される (ST250)。また、言語ごとにパターン辞書を参照して数詞解析によって得られたことを示す「数」も変数候補欄に格納される (ST250)。

【0064】

本実施例では、図 15 に示されるように単語 ID = 1-5、1-7、1-11、1-14 が数詞として抽出され、解析情報欄に「数詞」が格納され、変数候補欄に「NUM」と数詞解析による認定を示す「数」を格納される。

【0065】

ステップ ST250 の数詞解析の後、表記から固有名詞とわかるか否かが判定される (ST260)。この判定は、解析している文の言語によって決定される。たとえば、英語の場合、文頭でない単語で単語の先頭文字が大文字であれば、その単語は固有名詞であると決定可能である。このように、ステップ ST260 では、単語の形式から直ちに固有名詞であるか否かがわかるか否かが判定される。

【0066】

一方、たとえば、日本語、ドイツ語は単語の形式から直ちに固有名詞であるか否かがわからないので、この場合はステップ ST260 でわからないと判定される。

【0067】

ステップ ST260 で、表記から固有名詞であるか否かがわかる場合は、文頭でない単語で単語の先頭文字が大文字である単語があるか否か検索して、この条件に該当する単語があれば、それを固有名詞として認定し、図 15 の解析結果の解析情報欄に「固有名詞」を

格納する (S T 2 7 0)。

【0068】

さらに、この「固有名詞」を付された単語を変数候補として、変数候補欄に固有名詞を示す「PN」を格納する (S T 2 7 0)。さらにまた、この「固有名詞」を付された単語は、固有名詞辞書を参照して確定されたことを示すために変数候補欄に「辞」を格納する (S T 2 7 0)。

【0069】

本実施例では、ステップ S T 2 7 0 の固有名詞表記抽出では、図 1 5 の解析結果の対訳対 I D = 1 から、まだ変数候補が定まっていないものに対して、単語の先頭文字が大文字で、それ以外の文字が大文字でない単語列を調べると、該当する単語がないので何も実行しない。 10

【0070】

図 1 6 は、図 1 5 の処理後、参照するパターン集である。このパターン集は、メモリ 1 5 0 0 に格納されている。

ステップ S T 2 6 0 において表記から固有名詞とわからないと判定された場合、または、ステップ S T 2 7 0 の処理の後、上述した日本語の場合と同様に図 1 6 に示されるパターン集が参照されて、対訳対 I D = 1 の英語文のうち図 1 6 に示されるパターンに合う英語文中の単語があるかが調べられる (S T 2 8 0)。

ここで、p I D 欄、標準形欄、P 単語 I D 欄、単語欄、および変数候補欄は、上記の図 7 の対応する説明と同様である。 20

また、図 1 6 の p I D = p 4 には英語の場合での時刻のパターンが記載されているが、そのパターンによって行われる動作は、上記の図 7 の対応する説明と同様である。

【0071】

図 1 7 は、図 1 6 のパターン集にもとづいて抽出されたパターン情報である。このパターン情報は、メモリ 1 5 0 0 に格納される。

図 1 6 のパターン集を参照して、図 1 5 に示される解析結果から、パターン集のパターンに対応する先頭の単語 I D、パターン集のパターンに対応する末尾の単語 I D、その単語の標準形をその単語に対応させた具体的な形式に変換したもの、および、そのパターン集の中の変数候補欄に格納されている記号に対応している、解析結果の変数候補欄に含まれる単語 I D の 4 つがパターン情報として抽出される (S T 2 8 0)。 30

【0072】

ただし、パターンの標準形が 1 単語からなる場合は、パターン集のパターンに対応する先頭の単語 I D とパターン集のパターンに対応する末尾の単語 I D とは同一の単語 I D となるので、この場合は、パターン集のパターンに対応する末尾の単語 I D を省略してパターン情報として抽出される。

【0073】

本実施例では、図 1 5 の解析結果の対訳対 I D = 1 から、事前に準備している図 1 6 の「パターン集」の言語が「英」で、パターンにある単語や品詞が一致する箇所を調べる。パターン中の空白の項目は任意の単語や品詞と一致することを示す。

【0074】

図 1 6 の「パターン集」の p I D = p 3 に関しては、図 1 5 の解析結果の対訳対 I D = 1 の言語欄「英」において、単語 I D = 1 - 5、1 - 7、1 - 1 1、1 - 1 4 と一致する。また、パターンの変数候補欄が「NUM」となっているので、図 1 5 の解析結果の対訳対 I D = 1 の言語欄「英」の各単語 I D = 1 - 5、1 - 7、1 - 1 1、1 - 1 4 の変数候補欄に「NUM」を格納し、変数候補認定がパターンによるので「パ」を変数候補欄に格納する (図 1 8)。

【0075】

さらに、図 1 6 の「パターン情報」の対訳対 I D = 1 には、p I D = p 3 と一致した図 1 5 の解析結果の対訳対 I D = 1 の言語欄「英」の該当箇所の単語 I D = 1 - 5、1 - 7、1 - 1 1、1 - 1 4 と、パターンの標準形「6」、「4」、「1」、「4」と、パターン 50

中で変数候補となる単語 I D = 1 - 5、1 - 7、1 - 11、1 - 14 をそれぞれ変数候補欄に格納する。なお、図 7 のパターン集の標準形は表記と“\$ 番号”の並びであり、“\$ 番号”は P 単語 I D 欄と同じ番号で一致した単語（品詞が数詞の場合は半角のアラビア数字に変換されている）に置換される。

【0076】

図 19 は、品詞ごとに日本語とその日本語に対応する英語の単語をリスト化してある対訳辞書である。この対訳辞書は、メモリ 1500 に格納されている。図 19 に示されるように、対訳辞書は、同じ意味の日本語単語と英語単語とが品詞とともに関連づけられている。

図 18 の解析結果が得られた後、図 19 の対訳辞書を参照して、日本語の品詞と単語にもとづいて、その日本語の単語に対応する英語の単語を抽出する（ST310）。

10

【0077】

図 20 は、図 19 の対訳辞書にもとづいて、図 9 の日本語文に対する解析結果と図 18 の解析結果との単語 I D を対応付けた対応付け情報である。この対応付け情報は、メモリ 1500 に格納される。

【0078】

図 20 に示される対応付け情報は、ある対訳対 I D の文に対して、関連づけられた日本語の単語および英語単語の単語 I D を示す。対応付け情報には、対訳対 I D ごとに所定の対応付け手法（辞書、文字、パターン；文字、パターンについては後述）によって対応付けられた日本語単語および英語単語の単語 I D が格納される。また、対応付けられた両言語の単語 I D をひとまとまりとして、t I D という識別番号が付される。

20

【0079】

対応付けられた日本語単語および英語単語が抽出され、それらの単語 I D が図 20 に示されるように対応付け情報として各欄に格納される。さらに、日本語単語および英語単語の単語を対応付けた際に対訳辞書が参照されたことも対応付け手法欄に示される（ST310）。

【0080】

本実施例では、図 18 の解析結果の対訳対 I D = 1 から、日本語の品詞と単語をキーワードとして図 19 「対訳辞書（日英）」を検索し、得られた対訳と一致する英語の単語列を見つけると、図 20 の対応付け情報の対訳対 I D = 1 で、日本語単語 I D = 1 - 1 と英語単語 I D = 1 - 1 および、日本語 I D = 2 - 1 と英語単語 I D = 1 - 11 から 1 - 12 ままで該当する。その対応する日本語の単語と英語の単語列に関して、図 20 の対応付け情報の対訳対 I D = 1 に単語 I D（日）欄と単語 I D（英）欄に各単語 I D を格納し、対応付け手法欄に“辞書”を格納する。

30

【0081】

図 21 は、文字を全角に変換して、図 9 の日本語文に対する解析結果と図 18 の解析結果との単語 I D を対応付けた対応付け情報である。この対応付け情報は、メモリ 1500 に格納される。

図 18 の解析結果の対訳対 I D = 1 から、日本語の単語の単語を構成する文字と英語の単語を構成する文字を全角に変換したものを比較して、双方の単語の文字が一致するか否かを調べる（ST320）。双方の単語の文字が一致する双方の単語 I D を抽出して、図 20 のように対応付け情報に単語 I D を格納する（ST320）。

40

【0082】

また、双方の言語の単語の文字が一致した単語は、図 21 に示されるように、単語 I D（日）欄と単語 I D（英）欄に各単語 I D を格納し、対応付け手法欄に“文字”を格納する。

【0083】

本実施例では、日本語単語 I D = 2 - 2 と英語単語 I D = 1 - 7 および、日本語単語 I D = 4 - 1 と英語単語 I D = 1 - 5 および、日本語単語 I D = 4 - 2 と英語単語 I D = 1 - 6 および、日本語単語 I D = 4 - 3 と英語単語 I D = 1 - 7 が該当する。したがって、図

50

21の対応付け情報の対訳対ID=1、単語ID(日)欄と単語ID(英)欄に各単語IDを格納し、対応付け手法欄に“文字”を格納する。

【0084】

図9および図18の解析結果の対訳対ID=1から、日本語のカタカナ連続および読みが振られている固有名詞に対して、英語の読みと類似しているか否かを調べる(ST330)。たとえば、各言語の読みを音節切りした後、先頭音素の連続を求め、比較の際に言語の違いを補正して類似しているか否かを調べる。

【0085】

類似していると判定された場合は、単語ID(日)欄と単語ID(英)欄に各単語IDを格納し、対応付け手法欄に“音韻”を格納する。

10

【0086】

しかしながら、本実施例では日本語単語ID=1-1の読みと類似する英語単語の読みはなかったので、図7の「対応付け情報」の対訳対ID=1には何も付与されない。

【0087】

図22は、図8および図17のパターン情報にもとづいて、図9の日本語文に対する解析結果と図18英語文に対する解析結果との単語IDを対応付けた対応付け情報である。この対応付け情報は、メモリ1500に格納される。

図8および図17のパターン情報の対訳対ID=1から、両言語で標準形が一致するものを調べる(ST340)。双方言語の単語の標準形が一致した単語は、図22に示されるように、単語ID(日)欄と単語ID(英)欄に各単語IDを格納し、対応付け手法欄に“パターン(p1, p3)”を格納する。

20

【0088】

本実施例では、図8および図17のパターン情報の対訳対ID=1から、両言語で標準形が一致するものを調べると、日ID=1と英ID=2および、日ID=1と英ID=4および、日ID=2と英ID=1および、日ID=3と英ID=2および、日ID=3と英ID=4となる。したがって、情報欄のパターン一致単語の項の単語IDを図22の対応付け情報の対訳対ID=1の単語ID(日)欄と単語ID(英)欄に格納し、対応付け手法欄に“パターン(日本語のpID、英語のpID)”を格納する。

【0089】

たとえば、図22のtID=7の場合、日本語では、図7のpID=p1により、単語ID=2-2がpID=1で一致し、標準形の\$1に対応する。この単語は、pID=1の変数候補欄にNUMがあることから、変数候補でもある。英語では、図16のpID=p3により、単語ID=1-7がpID=1で一致し、標準形は\$1に対応する。この単語は、pID=1の変数候補欄にNUMがあることから、変数候補でもある。つぎに、両言語の標準形に関して対応を求めると、\$1(日本語)=\$1(英語)となるので、日本語の\$1に対する変数候補の単語ID=2-2と英語の\$1に対する変数候補の単語ID=1-7の対応が得られる。

30

【0090】

ステップST350の最適対応選択では、図22の対応付け情報の対訳対ID=1と以下の計算式を利用して対応度を計算し、対応度が最大となる単語対応の組み合わせを求める。なお、重みの値は確からしいものを大きな値として設定する。しかし、この重みの付け方は一例でありこれに限定されない。

40

対応度Tは、一方の言語の文中のある区間について計算される。ある区間とは、この例では日本語における文節であるが、句、節でもよい。

この区間内の各単語対応の対応度 $T_i(D)$ の i についての和で表される。Tも $T_i(D)$ もそれらの値を計算する際には単語を被計算対象として重複して計算しない。各言語とも単語対応が接続していればそれを1つの単語と同様な扱いで計算する。対応度T、および各単語対応の対応度 $T_i(D)$ は、次式で定義される。

【0091】

対応度 $T = \sum T_i(D)$

50

ここで、Dは、ある区間における、各言語の文で接続して並んでいる単語について、2つの言語間での単語対応を示している。たとえば、図9および図18の日本語文と英語文とでは、文節番号2の区間では、「前半」と「first half」の単語対応をD1として、「4」と「4」の単語対応がD2とすると、DはD1やD2が該当する。これらのすべての単語対応がついている区間ごとに $T_i(D)$ を計算して、それをすべての区間について和をとると、その和が対応度Tである。ここで、 $T_i(D)$ のiは、ある区間を示すサフィックスであり、一般に自然数である。

【0092】

単語対応の対応度

単語対応(D)の対応度 $T_i(D)$ は次式で定義される。ここで最初の項の和はswについてとり、つぎの項の和はtwについてとる。 10

【0093】

$$T_i(D) = (\sum (aw(sw) \times n(sw))) \times (\sum (aw(tw) \times n(tw)))$$

sw: 文で接続して並んでいる第1自然言語の単語、かつ単語対応のついている単語

tw: 単語連続に含まれる第2自然言語の単語、かつ単語対応のついている単語

aw(x): xの対応付け手法の重み

aw(x) = 1.0: xが辞書、パターンの場合

aw(x) = 0.9: xが文字の場合

aw(x) = 0.8: xが音韻の場合 20

n(x): xの単語数

ここで、n(x)は、形態素解析を実行する場合はその直後(ステップST220の直後)で単語IDに対応する単語(n(x)のxに対応する)について実行される。

また、形態素解析を実行しない場合、n(x)は、ステップST240の固有辞書検索直前の単語切りが実行されたのちの状態ですべての単語IDに対応する単語について実行される。たとえば、n(日本) = 1、n(Japan) = 1、n(United States) = 2となる。

【0094】

図22の対応付け情報の対訳対ID = 1から、日本語文節に限定して最適対応を求める。単語ID(日)で文節番号が同じ単語の組み合わせを求め、その日本語の単語と対応する英語の単語から対応度を計算する(ST351)。 30

【0095】

文節番号1では、図22の対応付け情報の対訳対ID = 1のtID = 1のみであるので以下のようなになる。

【0096】

$$\begin{aligned} \text{対応度} T &= (aw(\text{日本}) \times n(\text{日本})) \times (aw(\text{Japan}) \times n(\text{Japan})) \\ &= (1 \times 1) \times (1 \times 1) \\ &= 1 \end{aligned}$$

40

文節番号2では、図22の対応付け情報の対訳対ID = 1のtID = 2、3、7、8が対象となる。組み合わせは対応単語数と対応手法により複数作成される。すなわち、そのtIDの組み合わせは、{2}、{3}、{7}、{8}、{2, 3}、{2, 7}、{2, 8}である。この組み合わせのうち、tID = 2とtID = 7もしくはtID = 2とtID = 8で対応度は最大となる。

【0097】

tID = 2とtID = 7の場合の対応度を以下に示す。

$$\begin{aligned}
 \text{対応度}T &= (\text{aw}(\text{前半}) \times \text{n}(\text{前半})) \times (\text{aw}(\text{first}) \times \text{n}(\text{first}) + \\
 &\quad \text{aw}(\text{half}) \times \text{n}(\text{half})) + (\text{aw}(4) \times \text{n}(4)) \times (\text{aw}(4) \times \text{n}(4)) \\
 &= (1 \times 1) \times (1 \times 1 + 1 \times 1) + (0.9 \times 1) \times (0.9 \times 1) \\
 &= 2.81
 \end{aligned}$$

さらに、t I D = 2 と t I D = 8 の場合の対応度を以下に示す。

$$\begin{aligned}
 \text{対応度}T &= (\text{aw}(\text{前半}) \times \text{n}(\text{前半})) \times (\text{aw}(\text{first}) \times \text{n}(\text{first}) + \\
 &\quad \text{aw}(\text{half}) \times \text{n}(\text{half})) + (\text{aw}(4) \times \text{n}(4)) \times (\text{aw}(\text{four}) \times \text{n}(\text{four})) \\
 &= (1 \times 1) \times (1 \times 1 + 1 \times 1) + (0.9 \times 1) \times (0.9 \times 1) \\
 &= 2.81
 \end{aligned}$$

10

文節番号 3 では、該当するものがないので選択されない。

文節番号 4 では、図 2 2 の対応付け情報の対訳対 I D = 1 の t I D = 4、5、6、9、10、11 が対象となる。組み合わせは対応単語数と組み合わせ手法により複数個作成されるが、最大となるのは、t I D = 5 の対応手法が文字でありかつ、t I D = 9、10 の対応手法がパターンであって、日英ともに連続する場合である。

【0098】

$$\begin{aligned}
 \text{対応度}T &= (\text{aw}(6) \times \text{n}(6) + \text{aw}(-) \times \text{n}(-) + \text{aw}(4) \times \text{n}(4)) \times \\
 &\quad (\text{aw}(6) \times \text{n}(6) + \text{aw}(-) \times \text{n}(-) + \text{aw}(4) \times \text{n}(4)) \\
 &= ((1 \times 1) + (0.9 \times 1) + (1 \times 1)) \times \\
 &\quad ((1 \times 1) + (0.9 \times 1) + (1 \times 1)) \\
 &= 2.9 \times 2.9 \\
 &= 8.41
 \end{aligned}$$

20

この結果、図 2 3 の最適対応情報の対訳対 I D = 1 欄に対応付け情報の t I D の列 1、2、5、7、8、9、10 が格納される (S T 3 5 1)。

30

【0099】

図 2 3 は、図 2 2 に示されている t I D から対応度が最大となる t I D の組み合わせを示す最適対応情報である。この最適対応情報は、メモリ 1 5 0 0 に格納される。

最適対応情報の対訳対 I D = 1 の t I D = 1、2、5、7、8、9、10 から、両言語で重複のない組み合わせを求めて対応度を計算する (S T 3 5 2)。ここで、両言語で重複のない組み合わせとは、第 1 自然言語文と第 2 自然言語文のそれぞれで単語 I D が重複しない t I D の組み合わせということである。具体的には、たとえば、実施例 1 のステップ S T 3 5 2 で t I D = 1、2、5、7、8、9、10 である。逆に、両言語で重複のある組み合わせは、日本語では t I D = 7、8 であり、英語では t I D = 7、10 である。本実施例では、ステップ S T 3 5 2 の結果、対応度が最大となる t I D = 1、2、5、8、9、10 が選択される。この最大となる対応度 T は、

40

$$\begin{aligned}
 \text{対応度} T &= (aw(\text{日本}) \times n(\text{日本}) + aw(\text{Japan}) \times n(\text{Japan})) + \\
 &\quad (aw(\text{前半}) \times n(\text{前半}) + \\
 &\quad (aw(\text{first}) \times n(\text{first}) + aw(\text{half}) \times n(\text{half}))) + \\
 &\quad (aw(4) \times n(4) + aw(\text{four}) \times n(\text{four})) + \\
 &\quad (aw(6) \times n(6) + aw(-) \times n(-) + aw(4) \times n(4)) \times \\
 &\quad (aw(6) \times n(6) + aw(-) \times n(-) + aw(4) \times n(4)) \\
 &= 1 + 2.81 + 8.41 = 12.22
 \end{aligned}$$

10

となる。

【0100】

ステップ S T 3 0 0 の語句対応付けで得られた図 2 3 の「最適対応情報」の対訳対 I D = 1 にある「対応付け情報」の対訳対 I D = 1 の t I D から、テンプレートとしての確信度を計算する (S T 4 0 0)。確信度として文全体の内容語の対応割合とした場合の計算式を以下に示す。ここで、内容語とは、名詞、形容詞、動詞、および、副詞のようにもっぱら実質的な意味を担う単語のことである。

【0101】

確信度 1 (文全体の内容語の対応割合)

20

確信度 C = (第 1 自然言語と第 2 自然言語の内容語の対応語数) / (第 1 自然言語と第 2 自然言語の内容語の単語数)

本実施例では、日本語の内容語は、助詞を除いた単語 I D = 1 - 1、2 - 1、2 - 2、2 - 3、3 - 1、4 - 1、4 - 2、および、4 - 3 であるので、日本語での内容語の単語数は 8 である。一方、英語の内容語は、前置詞と冠詞を除いた単語 I D = 1 - 1、1 - 2、1 - 5、1 - 6、1 - 7、1 - 8、1 - 11、1 - 12、1 - 14、および、1 - 15 であるので、英語での内容語の単語数は 10 である。

他方、図 2 2 を参照すると、日本語の対応付けられている単語の単語 I D は、1 - 1、2 - 1、2 - 2、4 - 1、4 - 2、および、4 - 3 であるので、日本語の対応語数は 6 である。一方、同様に図 2 2 を参照すると、英語の対応付けられている単語の単語 I D は、1 - 1、1 - 5、1 - 6、1 - 7、1 - 11、1 - 12、および、1 - 14 であるので、英語の対応語数は 7 である。したがって、

30

$$\begin{aligned}
 \text{確信度 } C &= (6 + 7) / (8 + 10) \\
 &= 0.72
 \end{aligned}$$

となる。

【0102】

この確信度の値からテンプレート翻訳ルールを作成するか判定し、条件を満たせば、テンプレート翻訳ルールを作成し、出力する (S T 5 0 0)。たとえば、条件を確信度が閾値 0.7 以上でかつ最も高いものを残すとする。この条件において上記の計算された確信度は、確信度が 0.7 以上であるので、図 1 8 の解析結果の対訳対 I D = 1 と図 2 2 の「対応付け情報」の対訳対 I D = 1 と「最適対応情報」の対訳対 I D = 1 を用いて、図 8 の「テンプレート」の対訳対 I D = 1 が作成されることになる。

40

【0103】

図 2 4 は、図 5 の対訳集から作成されたテンプレートである。

日本語欄には、“(単語表記 品詞)”の列挙、英語欄は“単語表記”の列挙であり、変数箇所の対応情報欄には、“(日本語の変数記号 英語の変数記号 両言語で共通する品詞)”の列挙となる。

【0104】

このテンプレート翻訳ルールをテンプレート翻訳ルールデータベース 1 4 0 0 に格納する

50

(S T 6 0 0)。

【0 1 0 5】

図1のメモリ1500にある図24のテンプレート翻訳ルールをテンプレート翻訳ルールデータベース1400へ格納する。

【0 1 0 6】

図25は、本発明の実施例1において入力される対訳対IDが2の対訳集である。この対訳集は、メモリ1500に格納されている。図26は、対訳対ID=1と同様の処理を行って、図25に示されている日本語文および英語文に対して形態素解析を実行した結果、もしくは、図10から図18に対応する処理を実行した結果を反映する解析結果である。この解析結果は、メモリ1500に格納される。図27は、図26の解析結果との単語IDを対応付けた対応付け情報である。この対応付け情報は、メモリ1500に格納される。

10

本実施例の対訳対IDが1の場合と同様に、図2のステップS T 1 0 0の対訳対入力で、図25の「対訳集」から対訳対ID=2の対訳対を読み込む。なお、対訳対ID=2は、対訳対ID=1の日本語表記“前半”が“最初”に置き換わっていること以外是对訳対ID=1の場合と同様である。以下では対訳対ID=2の処理のうち対訳対ID=1と異なる処理を主に述べる。

【0 1 0 7】

ステップS T 2 0 0は対訳対ID=1と同様に処理される。

ステップS T 3 0 0の語句対応付けでは、図4のステップS T 3 1 0の対訳辞書対応において、図26の解析結果の対訳対ID=2の日本語単語ID=1-1と英語単語ID=1-1のみ該当する。そのため、ステップS T 3 5 0の最適対応選択で利用する図27の対応付け情報の対訳対ID=2と対訳対ID=1のt ID=2が削除されたものは同じ情報が収められている。

20

【0 1 0 8】

ステップS T 3 5 0の最適対応選択では、ステップS T 3 5 1の最適部分対応において、図27の「対応付け情報」の対訳対ID=2から最適対応を求める。

【0 1 0 9】

日本語文節ごとに対応度を求めると、図22の対訳対ID=1の対応付け情報と異なっているのは文節番号2でのみである。文節番号2では、図27の「対応付け情報」のt ID=2、6、7が対象となる。そのうち、t ID=6もしくはt ID=7となるものが最大となる。

30

【0 1 1 0】

t ID=6の場合の対応度を以下に示す。

【0 1 1 1】

$$\begin{aligned}\text{対応度}T &= (aw(4) \times n(4)) \times (aw(4) \times n(4)) \\ &= (1 \times 1) \times (1 \times 1) \\ &= 1\end{aligned}$$

40

t ID=7の場合の対応度を以下に示す。

【0 1 1 2】

$$\begin{aligned}\text{対応度}T &= (aw(4) \times n(4)) \times (aw(\text{four}) \times n(\text{four})) \\ &= (1 \times 1) \times (1 \times 1) \\ &= 1\end{aligned}$$

この結果、図7の「最適対応情報」の対訳対ID=2欄に対応付け情報のt ID=1、4、6、7、8、9が格納される(S T 3 5 1)。

【0 1 1 3】

50

図 28 は、図 27 に示されている t I D から対応度が最大となる t I D の組み合わせを示す最適対応情報である。この最適対応情報は、メモリ 1500 に格納される。

「最適対応情報」の対訳対 I D = 2 の t I D = 1、4、6、7、8、9 から、両言語で重複のない組み合わせを求めて対応度を計算した結果、対応度が最大となる t I D = 1、4、7、8、9 が選択される (S T 3 5 2)。

【0 1 1 4】

この最大となる対応度 T は、

$$\begin{aligned} \text{対応度 } T &= (\text{aw}(\text{日本}) \times n(\text{日本}) + \text{aw}(\text{Japan}) \times n(\text{Japan})) + \\ &\quad (\text{aw}(4) \times n(4) + \text{aw}(\text{four}) \times n(\text{four})) + \\ &\quad (\text{aw}(6) \times n(6) + \text{aw}(-) \times n(-) + \text{aw}(4) \times n(4)) \times \\ &\quad (\text{aw}(6) \times n(6) + \text{aw}(-) \times n(-) + \text{aw}(4) \times n(4)) \\ &= 1 + 1 + 8.41 = 10.41 \end{aligned}$$

10

となる。

【0 1 1 5】

ステップ S T 3 0 0 の語句対応付けで得られた図 28 の「最適対応情報」の対訳対 I D = 2 にある「対応付け情報」の対訳対 I D = 2 の t I D から、テンプレートとしての確信度を計算する (S T 4 0 0)。確信度として文全体の内容語の対応割合とした場合の計算式を以下に示す。

20

【0 1 1 6】

確信度 1 (文全体の内容語の対応割合)

$C = (\text{第 1 自然言語と第 2 自然言語の内容語の対応語数}) / (\text{第 1 自然言語と第 2 自然言語の内容語の単語数})$

この例では、上記の図 23 に関連して説明したように単語数は日本語が 8、英語が 10、日本語の対応語数が 5、英語が 5 である。したがって、

$$\begin{aligned} \text{確信度 } C &= (5 + 5) / (8 + 10) \\ &= 0.55 \end{aligned}$$

30

となる。

【0 1 1 7】

ステップ S T 5 0 0 のテンプレート翻訳ルール作成では、その確信度の値が閾値以上であれば、テンプレート翻訳ルールを作成して、出力する。たとえば、上述した例と同様に確信度が閾値 (0.7) 以上でかつ最も高いものを残すとする、上記で計算した値は、確信度が 0.7 未満であるので、テンプレートは作成されない。

【0 1 1 8】

以上より、対訳間の単語対応割合により、テンプレート翻訳に適したルールを選別できることが示された。

【0 1 1 9】

40

(実施例 2)

本実施例では、変数箇所の認定手法や対応付け手法の確からしさによって、テンプレート翻訳に適したルールの選別ができることを示す。

【0 1 2 0】

図 29 は、本発明の実施例 2 において入力される対訳対 I D が 1 の対訳集である。この対訳集は、メモリ 1500 に格納されている。

図 1 のインタフェース部 1100 では、図 29 の対訳集が入力され、図 1 のメモリ 1500 に格納される (図 2 の S T 1 0 0)。図 1 のメモリ 1500 にある図 29 の対訳集から各対訳対を対訳対 I D 順に図 2 の S T 2 0 0 から S T 5 0 0 の処理を行う。以下、図 2 に沿って説明する。

50

【0121】

図29の「対訳集」の対訳対ID=1を選ぶ。ステップST200の変数候補認定で、読み込んだ対訳対について、言語ごとに図3の処理を行う。

【0122】

以下、図3に沿って、最初に日本語の場合を説明し、次に中国語の場合を説明する。日本語の処理において、形態素解析手段が用意されていたとすると、図3のステップST210において「できる」が選択される。

【0123】

図30は、本発明の実施例2において日本語文に対して形態素解析(ST220)を実行した場合の解析結果である。この解析結果は、メモリ1500に格納される。

10

図29の対訳集の対訳対ID=1の言語欄“日”の文を形態素解析した結果、図30の解析結果の対訳対ID=1の言語欄“日”の単語欄に各単語表記を格納し、単語ID欄に“文節番号－文節内の単語番号”を格納し、解析情報欄に品詞と読みを格納する。

【0124】

さらに、上記の結果から単語ID=1-3、1-5の解析情報欄に固有名詞とあるので、変数候補欄に固有名詞を示す“PN”を格納し、形態素解析で変数候補が認定されたので、認定手法を示す“形”を格納する。

【0125】

図31は、図30の処理後、参照するパターン集である。このパターン集は、メモリ1500に格納されている。図32は、図31のパターン集にもとづいて抽出されたパターン情報である。このパターン情報は、メモリ1500に格納される。

20

パターンマッチでは、図30の解析結果から、事前に準備している図31のパターン集の言語欄が“日”となるパターンを用いて、パターンと合う日本語文中の単語や品詞を調べる(ST280)。なお、パターン中の空白の項目は任意の単語や品詞との一致することを示す。

しかしながら、本実施例では該当するものがないので、何も実行しない。

【0126】

一方、中国語の処理において、形態素解析手段が用意されていないとすると、図3のステップST210において「できない」が選択される。

図33は、中国語の漢字と読みを対応付けてある音韻辞書である。この音韻辞書は、メモリ1500に格納されている。図34は、本発明の実施例2において中国語文に対して単語IDを付してその単語の読み付した解析結果である。この解析結果は、メモリ1500に格納される。

30

単語切りでは、通常中国語は空白を単語境界としないので、仮に1文字を1単語とみなして、図29の「対訳集」の対訳対ID=1の言語欄“中”の文に対し、単語切りを実行する(ST230)。図30の解析結果の対訳対ID=1の言語欄“中”の単語欄に単語を、単語ID欄に“1－単語番号”を格納する。また、図33の「音韻辞書(中)」を用いて、読み欄に読みを格納する。

【0127】

図35は、中国語の固有名詞をリスト化してある固有名詞辞書である。この固有名詞辞書は、メモリ1500に格納されている。図36は、図35の固有名詞辞書を参照して固有名詞を解析情報に付してさらに変数候補を抽出した解析結果である。この解析結果は、メモリ1500に格納される。

40

ステップST240の固有名詞辞書検索では、図34の解析結果の対訳対ID=1の言語欄“中”に対し、任意の単語列が固有名詞辞書に載っているか否かを調べ、中国語文に対して最長の単語列を優先して選択する。すると、単語ID=1-5から1-6が一致し、また単語ID=1-9から1-11が一致するので、その単語列を1語にまとめて単語IDを振りなおす(図36の解析結果)。図36の解析結果において解析情報欄に“固有名詞”を格納し、変数候補欄に“PN”と辞書による認定を示す“辞”を格納する。

【0128】

50

ステップ S T 2 5 0 の数詞解析では、図 3 6 の解析結果の対訳対 I D = 1 から、まだ変数候補となっていない語に対して、各単語列から数詞となるものを調べる。しかし、本実施例では該当するものがないので何も実行しない。

ステップ S T 2 6 0 では、中国語は表記から固有名詞とわからないので、「わからない」を選択する。

【 0 1 2 9 】

図 3 7 は、図 3 6 の処理後、参照するパターン集である。このパターン集は、メモリ 1 5 0 0 に格納されている。

パターンマッチでは、図 3 6 の解析結果の対訳対 I D = 1 から、事前に準備している図 3 7 の「パターン集」の言語が“中”で、パターンにある単語や品詞が一致する箇所を調べる (S T 2 8 0)。パターン中の空白の項目は、任意の単語や品詞と一致することを示す。しかしながら、本実施例では該当するものがないので、何も実行しない。

10

【 0 1 3 0 】

図 3 8 は、品詞ごとに日本語とその日本語に対応する中国語の単語をリスト化してある対訳辞書である。この対訳辞書は、メモリ 1 5 0 0 に格納されている。図 3 9 は、図 3 8 の対訳辞書にもとづいて、図 3 0 の日本語文に対する解析結果と図 3 6 の解析結果との単語 I D を対応付けた対応付け情報である。この対応付け情報は、メモリ 1 5 0 0 に格納される。

ステップ S T 3 1 0 の対訳辞書対応では、図 3 6 の解析結果の対訳対 I D = 1 から、日本語の品詞と単語をキーワードにして、図 9 の「対訳辞書 (日中)」を検索する。その結果、抽出された対訳と一致する中国語の単語列を見つけると、図 3 6 の解析結果の対訳対 I D = 1 で日本語単語 I D = 1 - 1 と中国語単語 I D = 1 - 1 から 1 - 2 および、日本語単語 I D = 1 - 2 と中国語単語 I D = 1 - 3 から 1 - 4 および、日本語単語 I D = 1 - 3 と中国語単語 I D = 1 - 5 および、日本語単語 I D = 1 - 5 と中国語単語 I D = 1 - 8 が該当する。その対応する日本語の単語と中国語の単語列に関して、図 3 9 の対応付け情報の対訳対 I D = 1 の単語 I D (日) と単語 I D (中) に各単語 I D を格納し、対応付け手法欄に“辞書”を格納する。

20

【 0 1 3 1 】

ステップ S T 3 2 0 の文字対応では、図 3 0 および図 3 6 の解析結果の対訳対 I D = 1 から、日本語の単語について、単語を構成する文字列と中国語の単語を構成する文字とが一致するものを調べる。しかし、本実施例では、日本語の単語と同じ文字列は見つからないので何も実行しない。

30

【 0 1 3 2 】

図 4 0 は、音韻辞書にもとづいて、図 3 0 の日本語文に対する解析結果と図 3 6 の解析結果との単語 I D を対応付けた対応付け情報である。この対応付け情報は、メモリ 1 5 0 0 に格納される。

ステップ S T 3 3 0 の類似音韻対応では、図 3 0 および図 3 6 の解析結果の対訳対 I D = 1 から、日本語のカタカナ連続および読みが振られている固有名詞に対して読みを示す文字列が両言語で類似しているか否かを調べる。

類似しているか否かを調べるために、言語ごとに読み補正辞書が用意されている。読み補正辞書は、第 1 および第 2 言語の読み補正情報として先頭音素列とその補正音素列との対応関係を記憶している。この読み補正辞書によって、対応のつかない先頭音素は、削除可能であるとみなして、削除する。

40

【 0 1 3 3 】

そして、第 1 および第 2 言語の各単語の読み情報を各々対応する補正された音素列を決定し、各補正音素列で一致する音素数を算出する。この一致した音素数のその単語全体の音素数に対する割合を算出する。この割合が所定の閾値を越えていれば、それらの単語の読みは類似しているとして、それらの単語は対応付けられる (S T 3 3 0)。

【 0 1 3 4 】

たとえば、各言語の読みを音節切りした後、先頭音素の連続を求めて、比較の際に言語の

50

違いを補正して類似しているか否かを調べる。この場合、日本語単語 I D = 1 - 3 の読み “b u r a z i r u” は先頭音素列 “B - R - Z - R” となる。一方、中国語単語 I D = 1 - 5 の読み “b a x i” は先頭音素列 “B - X” となる。両者を読み補正辞書によって補正して比較すると、“B - _ - Z - _” (“_” は削除を意味する) と “B - Z” で一致する音素数の割合は 0.5 となる。

また、中国語単語 I D = 1 - 8 の読み “w u l a g u i” は先頭音素列 “W - L - G” となる。両者を補正して比較すると、“B - R - Z - _” と “B - R - Z” となり、一致する音素数の割合は 1 となる。したがって、一致する音素数の割合の大きい対応である日本語単語 I D = 1 - 3 と中国語単語 I D = 1 - 8 が得られる (S T 3 3 0)。しかし、この場合の日本語単語 I D = 1 - 3 と中国語単語 I D = 1 - 8 の対応付けは誤りである。 10

図 4 0 の「対応付け情報」の対訳対 I D = 1 の単語 I D (日) と単語 I D (中) に各単語 I D を格納し、対応付け手法欄に “音韻” を格納する。

ステップ S T 3 4 0 のパターン対応では、図 3 2 のパターン情報の対訳対 I D = 1 には、何も載っていないので、本実施例では何も実行しない。

【0135】

ステップ S T 3 5 0 の最適対応選択では、図 4 0 の対応付け情報の対訳対 I D = 1 と実施例 1 と同じ計算式を利用して対応度を計算し、対応度が最大となる単語対応の組み合わせを求める。

【0136】

図 4 1 は、図 4 0 に示されている t I D から対応度が最大となる t I D の組み合わせを示す最適対応情報である。この最適対応情報は、メモリ 1 5 0 0 に格納される。 20

ステップ S T 3 5 1 の最適部分対応では、図 4 0 の対応付け情報の対訳対 I D = 1 から、日本語文節に限定して最適対応を求める。単語 I D (日) で文節番号が同じ単語の組み合わせを求め、その日本語の単語と対応する中国語の単語から対応度を計算する。

【0137】

文節番号 1 では、図 4 0 の「対応付け情報」の全てが対象となる。組み合わせは対応単語数と対応手法により複数作成されるが、t I D = 1、2、3、4 となるものの下記の式 1 に示される対応度 T が最大となり、最適対応情報にこれらの t I D を格納する。

【0138】

【式 1】

$$\begin{aligned}
 \text{対応度 } T &= (\text{aw}(\text{サッカー}) \times n(\text{サッカー}) + \text{aw}(\text{決勝}) \times n(\text{決勝}) + \\
 &\quad \text{aw}(\text{ブラジル}) \times n(\text{ブラジル})) \times \\
 &\quad ((\text{aw}(\text{足球}) \times n(\text{足球}) + \text{aw}(\text{決賽}) \times n(\text{決賽}) + \text{aw}(\text{巴西}) \times n(\text{巴西})) + \\
 &\quad (\text{aw}(\text{ウルグアイ}) \times n(\text{ウルグアイ}))) \times (\text{aw}(\text{乌拉圭}) \times n(\text{乌拉圭})) \\
 &= ((1 \times 1) + (1 \times 1) + (1 \times 1)) \times ((1 \times 1) + (1 \times 1) + (1 \times 1)) + \\
 &\quad (1 \times 1) \times (1 \times 1) \\
 &= 9 + 1 \\
 &= 10
 \end{aligned}$$

【0139】

ステップ S T 3 5 2 の最適文対応では、図 4 0 の「対応付け情報」の対応のうちステップ S T 3 5 1 で得られた I D = 1、2、3、4 から、日本語の単語 I D で単語の組み合わせを求め、求められた日本語の単語と対応する中国語の単語から対応度を計算する。日本語は文節番号 1 が 1 文となるので、ステップ S T 3 5 1 で求めたものと同じ結果になる。

【0140】

図 2 のステップ S T 4 0 0 の確信度計算では、ステップ S T 3 0 0 の語句対応付けで得られた図 1 1 の「最適対応情報」の対訳対 I D = 1 にある「対応付け情報」の対訳対 I D = 1 の t I D から、テンプレートとしての確信度を計算する。 50

【0141】

確信度として、変数箇所の認定手法や対応付け手法を用いて定める。なお、各種重みの値は確からしいものを大きな値としているが、一例でありこれに限定されない。一般に確信度Cは、変数箇所の品詞、認定手法、対応付け手法が確からしく、かつ、変数箇所間で未対応となるものが少ないほど大きな値となる。

【0142】

確信度2（変数箇所の認定手法や対応付け手法の確からしさ）ここで、左側のΣは各言語についての和をとることを示し、その右側にあるΣは変数箇所に対応のある単語についての和をとることを示す。

【0143】

$C = (\sum \sum (w_1 \times w_2 \times w_3)) / (\text{両言語の変数箇所の数})$

重み（値は例）

w1：品詞の重み

w1 = 1：品詞が固有名詞または数詞の場合

w1 = 0.8：その他

w2：変数箇所の認定手法の重み

w2 = 1：認定手法が形態素解析の品詞、固有名詞辞書、パターンの場合

w2 = 0.9：認定手法が語頭大文字語の場合（英語等に限定）

w2 = 0.8：その他

w3：変数箇所の対応付け手法の重み

w3 = 1：対応付け手法が対訳辞書、パターンの標準形の場合

w3 = 0.9：対応付け手法が文字の場合

w3 = 0.8：対応付け手法が音韻の場合

w3 = 0.5：その他

各言語においては、変数箇所に対応のある単語が被計算対象として選択される。この例では、確信度は以下ようになる。

【0144】

【式2】

$$\begin{aligned} C &= (\text{確信度(ブラジル)} + \text{確信度(ウルグアイ)} + \\ &\quad \text{確信度(巴西)} + \text{確信度(乌拉圭)}) / (2+2) \\ &= ((1 \times 1 \times 1) + (1 \times 1 \times 1) + (1 \times 1 \times 1) + (1 \times 1 \times 1)) / 4 \\ &= 1 \end{aligned}$$

【0145】

ここで、確信度(X) = 単語Xについてw1 × w2 × w3の値とする。

【0146】

図42は、図29の対訳集から作成されたテンプレートである。

ステップST500のテンプレート翻訳ルール作成では、その確信度の値からテンプレート翻訳ルールを作成するか判定して、条件を満たせばテンプレート翻訳ルールを作成し、出力する。

【0147】

たとえば、確信度が閾値(0.7)以上でかつ最も高いものを条件として選択する。上記の計算した確信度は0.7以上であるので、図30および図36の解析結果の対訳対ID = 1と図40の対応付け情報の対訳対ID = 1と図41の最適対応情報の対訳対ID = 1を用いて、図42の「テンプレート」の対訳対ID = 1が作成される。

【0148】

なお、日本語欄には、“(単語表記 品詞)”の列挙、中国語欄は“単語表記”の列挙であり、変数箇所の対応情報欄には、“(日本語の変数記号 中国語の変数記号 両言語で共通する品詞)”の列挙となる。

10

20

30

40

50

【0149】

このテンプレート翻訳ルールをテンプレート翻訳ルールデータベース1400に格納する(ST600)。

【0150】

図43は、図35の固有名詞辞書に固有名詞が載っていなかった場合の、図40に対応する対応付け情報である。この対応付け情報は、メモリ1500に格納される。

上記処理で、仮に、図38の対訳辞書(日中)に固有名詞の対訳が載っていなかった場合について述べる。なお、本実施例の上記の例と異なる箇所について主に述べる。

【0151】

図4のステップST310の対訳辞書対応において、固有名詞の対応は得られないので、辞書による対応付けの結果は図43の対応付け情報となる。 10

【0152】

ステップST350の最適対応選択では、図43の対応付け情報の対訳対ID=1と実施例1と同じ計算式を利用して対応度を計算し、対応度が最大となる単語対応の組み合わせを求める。

【0153】

図44は、図43に示されているtIDから対応度が最大となるtIDの組み合わせを示す最適対応情報である。この最適対応情報は、メモリ1500に格納される。

ステップST351の最適部分対応では、図43の対応付け情報の対訳対ID=1から、日本語文節に限定して最適対応を求める。単語ID(日)で文節番号が同じ単語の組み合わせを求め、その日本語の単語と対応する中国語の単語から対応度を計算する。 20

【0154】

文節番号1では、図43の対応付け情報の全てが対象となる。組み合わせは対応単語数と対応手法により複数作成されるが、ID=1、2、3となるものが最大となり、最適対応情報に記録される。この例では、確信度は以下のようになる。

【0155】

【式3】

$$\begin{aligned} \text{対応度 } T &= (\text{aw(サッカー)} \times n(\text{サッカー}) + \text{aw(決勝)} \times n(\text{決勝})) \times \\ &\quad (\text{aw(足球)} \times n(\text{足球}) + (\text{决赛}) \times n(\text{决赛})) + \\ &\quad (\text{aw(ブラジル)} \times n(\text{ブラジル})) \times (\text{aw(乌拉圭)} \times n(\text{乌拉圭})) \\ &= ((1 \times 1) + (1 \times 1)) \times ((1 \times 1) + (1 \times 1)) + \\ &\quad (0.8 \times 1) \times (0.8 \times 1) \\ &= 4.64 \end{aligned}$$

30

【0156】

ステップST352の最適文対応では、図43の対応付け情報の対応のうちステップST351で残ったID=1、2、3から、日IDで単語の組み合わせを求める。求められた日本語の単語と対応する中国語の単語から対応度を計算すると、日本語は1文節が1文となっているので、ステップST351で求めたものと同じ結果になる。 40

【0157】

図2のステップST400の確信度計算では、ステップST300の語句対応付けで得られた図11の「最適対応情報」の対訳対ID=1にある「対応付け情報」の対訳対ID=1のtIDから、テンプレートとしての確信度を計算する。確信度は、以下のように計算される。

【0158】

【式4】

$$\begin{aligned}
 C &= (\text{確信度(ブラジル)} + \text{確信度(ウラ圭)}) / (2+2) \\
 &= ((1 \times 1 \times 0.8) + (1 \times 1 \times 0.8)) / 4 \\
 &= 1.6 / 4 \\
 &= 0.4
 \end{aligned}$$

【0159】

ステップ S T 5 0 0 のテンプレート翻訳ルール作成では、その確信度の値からテンプレート翻訳ルールを作成するか判定して、条件を満たせば作成し、出力する。たとえば、確信度が閾値（0.7）以上でかつ最も高いものを条件として選択すると、上記の計算した結果は、確信度が0.7未満であるので、テンプレートは作成されない。10

以上より、変数箇所の認定手法や対応付け手法の確からしさによって、テンプレート翻訳に適したルールの選別ができることが示された。

【0160】

この例では第1自然言語を日本語、第2自然言語を英語または中国語とするが、これらの言語に限定されない。

この発明は、上述した実施の形態に限定されるものではなく、その技術的範囲において種々変形して実施することができる。

【0161】

【発明の効果】

本発明の翻訳ルール作成方法、装置およびプログラムによれば、テンプレート翻訳に利用可能なルールか判定するために、変数候補の認定手法や対応付け手法や文全体の単語対応割合から確信度を計算して、確信度の高いものを適切なテンプレート翻訳ルールとすることによって、信頼性の高いテンプレート翻訳用ルールを作成することが可能になる。20

【図面の簡単な説明】

【図1】本発明の実施形態に係る翻訳ルール作成装置の機能ブロック図。

【図2】本発明の実施形態に係る翻訳ルール作成装置で使用される翻訳ルール作成方法における処理の流れを示すフローチャート。

【図3】図2のステップ S T 2 0 0 の処理の流れを示すフローチャート。

【図4】図3のステップ S T 3 0 0 の処理の流れを示すフローチャート。30

【図5】本発明の実施例1において入力される対訳対 I D が1の対訳集。

【図6】本発明の実施例1において日本語文に対して形態素解析を実行した場合の解析結果。

【図7】図6の処理後、参照されるパターン集。

【図8】図7のパターン集にもとづいて抽出されたパターン情報。

【図9】図8のパターン情報を反映させた解析結果。

【図10】本発明の実施例1において英語文に対して単語 I D を付してその単語の読み付した解析結果。

【図11】英語単語とその品詞（前置詞、冠詞等）を対応づけてある付属語辞書。

【図12】図11の付属語辞書を参照して前置詞および冠詞を解析情報に付した解析結果。40

【図13】英語の固有名詞をリスト化してある固有名詞辞書。

【図14】図13の固有名詞辞書を参照して固有名詞を解析情報に付してさらに変数候補を抽出した解析結果。

【図15】数字辞書およびパターン辞書を参照して数詞を解析情報に付してさらに変数候補を抽出した解析結果。

【図16】図15の処理後、参照するパターン集。

【図17】図16のパターン集にもとづいて抽出されたパターン情報。

【図18】図17のパターン情報を反映させた解析結果。

【図19】品詞ごとに日本語とその日本語に対応する英語の単語をリスト化してある対訳 50

辞書。

【図 20】図 19 の対訳辞書にもとづいて、図 9 の日本語文に対する解析結果と図 18 の解析結果との単語 I D を対応付けた対応付け情報。

【図 21】図 20 に、文字を全角に変換して、図 9 の日本語文に対する解析結果と図 18 の解析結果との単語 I D を対応付けたものを付加した対応付け情報。

【図 22】図 21 に、図 8 および図 17 のパターン情報にもとづいて、図 9 の日本語文に対する解析結果と図 18 の解析結果との単語 I D を対応付けたものを付加した対応付け情報。

【図 23】図 22 に示されている t I D から対応度が最大となる t I D の組み合わせを示す最適対応情報。

10

【図 24】図 5 の対訳集から作成されたテンプレート。

【図 25】本発明の実施例 1 において入力される対訳対 I D が 2 の対訳集。

【図 26】図 25 に示されている日本語文および英語文に対して形態素解析を実行した結果、および、対訳対 I D が 1 での図 10 から図 19 に対応する処理と同様の処理を実行した結果を反映する解析結果。

【図 27】図 26 の解析結果との単語 I D を対応付けた対応付け情報。

【図 28】図 27 に示されている t I D から対応度が最大となる t I D の組み合わせを示す最適対応情報。

【図 29】本発明の実施例 2 において入力される対訳対 I D が 1 の対訳集。

【図 30】本発明の実施例 2 において日本語文に対して形態素解析を実行した場合の解析結果。

20

【図 31】図 30 の処理後、参照するパターン集。

【図 32】図 31、図 37 のパターン集にもとづいて抽出されたパターン情報。

【図 33】中国語の漢字と読みを対応付けてある音韻辞書。

【図 34】本発明の実施例 2 において中国語文に対して単語 I D を付してその単語の読み付した解析結果。

【図 35】中国語の固有名詞をリスト化してある固有名詞辞書。

【図 36】図 35 の固有名詞辞書を参照して固有名詞を解析情報に付してさらに変数候補を抽出した解析結果。

【図 37】図 36 の処理後、参照するパターン集。

30

【図 38】品詞ごとに日本語とその日本語に対応する中国語の単語をリスト化してある対訳辞書。

【図 39】図 38 の対訳辞書にもとづいて、図 31 の日本語文に対する解析結果と図 35 の解析結果との単語 I D を対応付けた対応付け情報。

【図 40】図 39 に、音韻辞書にもとづいて、図 30 の日本語文に対する解析結果と図 34 の解析結果との単語 I D を対応付けたものを付加した対応付け情報。

【図 41】図 40 に示されている t I D から対応度が最大となる t I D の組み合わせを示す最適対応情報。

【図 42】図 29 の対訳集から作成されたテンプレート。

【図 43】図 35 の固有名詞辞書に固有名詞が載っていなかった場合の、図 40 に対応する対応付け情報。

40

【図 44】図 43 に示されている t I D から対応度が最大となる t I D の組み合わせを示す最適対応情報。

【符号の説明】

1 0 0 0	翻訳ルール作成装置
1 1 0 0	インタフェース部
1 2 0 0	テンプレート翻訳ルール作成部
1 3 0 0	テンプレート翻訳ルール作成制御部
1 4 0 0	テンプレート翻訳ルールデータベース
1 5 0 0	メモリ

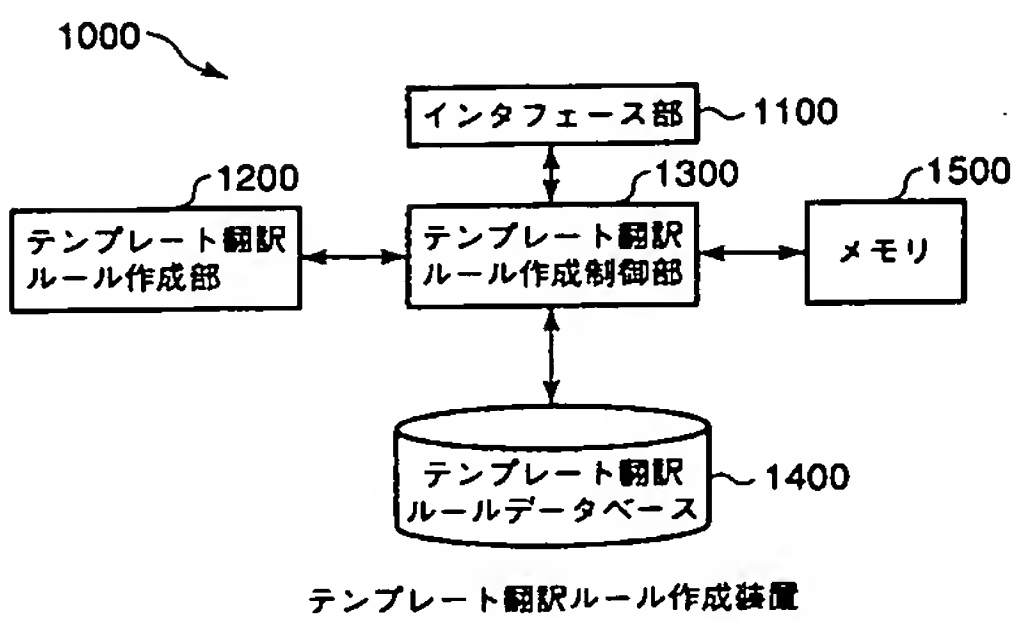
50

S T 1 0 0	対訳対入力
S T 2 0 0	変数候補認定
S T 3 0 0	語句対応付け
S T 4 0 0	確信度計算
S T 5 0 0	テンプレート翻訳ルール作成
S T 6 0 0	テンプレート翻訳ルール格納
S T 2 1 0	形態素解析が利用できるか否かの判定
S T 2 2 0	形態素解析
S T 2 3 0	単語切り
S T 2 4 0	固有名詞辞書検索
S T 2 5 0	数詞解析
S T 2 6 0	表記から固有名詞と分かるか否かの判定
S T 2 7 0	固有名詞表記抽出
S T 2 8 0	パターンマッチ
S T 3 1 0	対訳辞書対応
S T 3 2 0	文字対応
S T 3 3 0	類似音韻対応
S T 3 4 0	パターン対応
S T 3 5 0	最適対応選択
S T 3 5 1	最適部分対応
S T 3 5 2	最適文対応

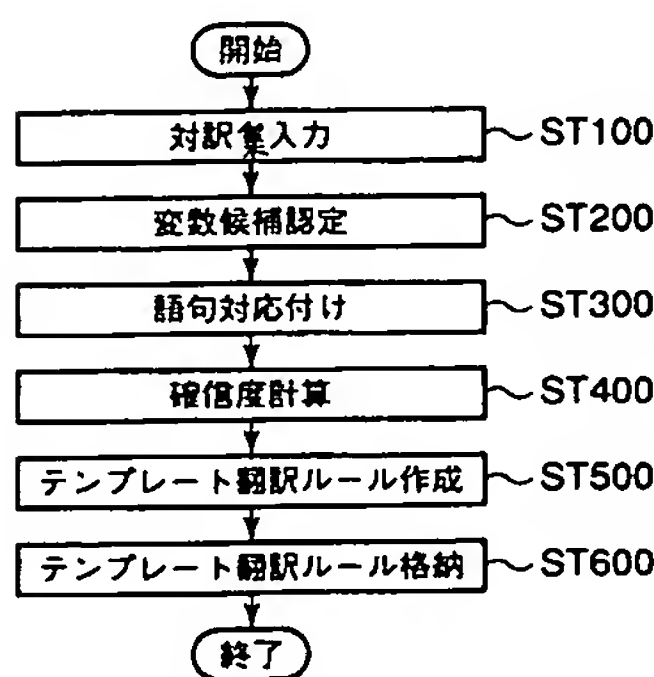
10

20

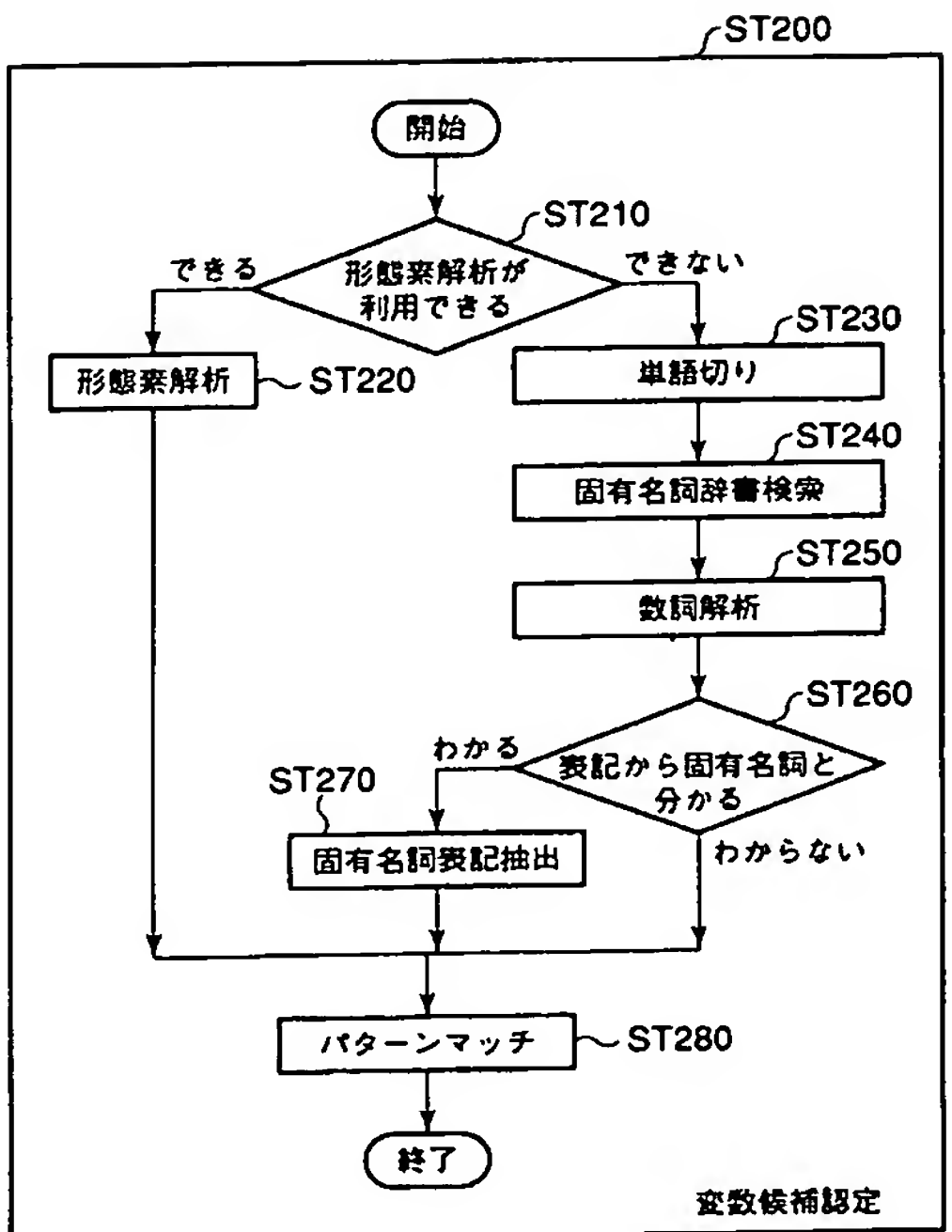
【 図 1 】



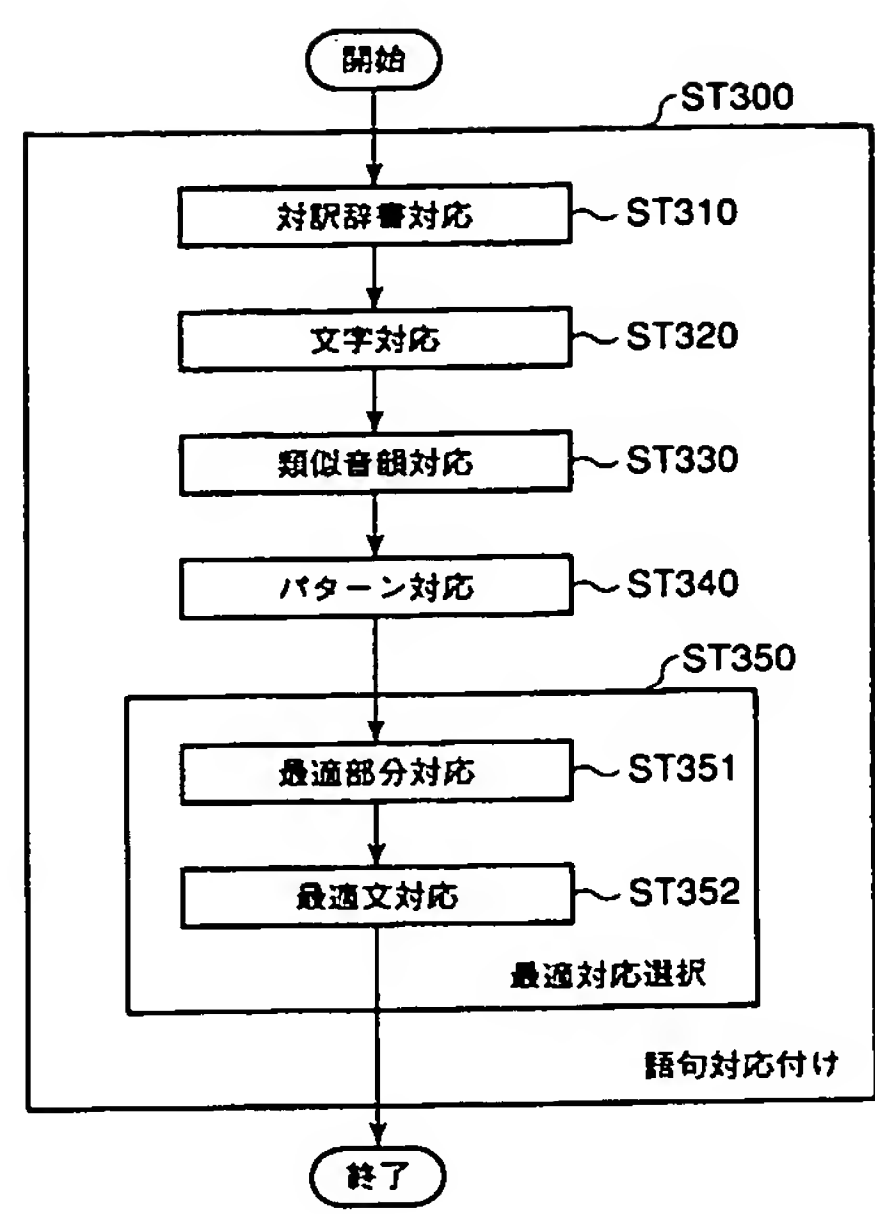
【 図 2 】



【 図 3 】



【図 4】



【図 5】

対訳集	対訳対ID	言語	文
	1	日	日本は前半4点を奪い6-4
		英	Japan cruised to a 6-4 advantage in the first half with four goals

【図 6】

	1-1	日本	固有名称	nippon	PN:形
	1-2	は	助詞		
	2-1	前半	名詞		
	2-2	4	数詞		NUM:形
	2-3	点	接尾		
	2-4	を	助詞		
	3-1	奪い	動詞		
	4-1	6	数詞		NUM:形
	4-2	.	記号		
	4-3	4	数詞		NUM:形
単語ID(文節-単語)					
単語					
解析情報					
読み					
変数候補					
範囲	日				
対訳対ID	1				

【図 7】

パターン集

pID	言語	標準形	パターン				
p1	日	\$1	P単語ID	1			
			単語				
			品詞	数詞			
			変数候補	NUM			
p2	日	\$1時\$3分	P単語ID	1	2	3	4
			単語		時		分
			品詞	数詞	接尾	数詞	接尾
			変数候補	NUM		NUM	

【図 8】

パターン情報(パターン一致単語:標準形:変数候補単語)					
対訳対ID	言語				
1	日	ID	1	2	3
		情報	(2-2:4:2-2)	(4-1:6:4-1)	(4-3:4:4-3)

【 図 9 】

言語ID	言語												
1	日	単語ID(文節-単語)	1-1	1-2	2-1	2-2	2-3	2-4	3-1	4-1	4-2	4-3	NUM : 形、バ
		単語	日本	は	前半	4	点	を	算い	6	-	4	
		解析情報	固有名詞	助詞	名詞	数詞	接尾	助詞	動詞	数詞	記号	数詞	
		読み	nippon										
		変数候補	PN : 形							NUM : 形、バ			

【 図 1 0 】

[illegible]

【 図 1 1 】

単語	品詞
to	前置詞
a	冠詞
in	前置詞
the	冠詞
with	前置詞
...	...

【图 1 2】

[illegible]

【 ㊦ 1 3 】

固有名詞辭書(英)

Japan

【 図 1 4 】

[illegible]

【 図 1 5 】

英 国																
词根对应 ID																
	单词ID (1-单词)	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12	1-13	1-14	1-15
	单词	Japan	cruised	to	a	6	.	4	advan- tage	in	the	first	half	with	four	goals
	解析情报 読み	固有名词 japan	cruised	前置词 to	冠词 a	数词 6		数词 4	advan- tago	前置词 in	冠词 the	数词 first	half	前置词 with	数词 four	goals
	変数候補	PN : 辞				NUM : 数		NUM : 数				NUM : 数			NUM : 数	

【 図 1 6 】

パターン集

pID	言語	標準形	パターン				
p3	英	\$1	P単語ID	1			
			単語				
			品詞	数詞			
			変数候補	NUM			
p4	英	\$1時\$3分	P単語ID	1	2	3	
			単語		:		
			品詞	数詞		数詞	
			変数候補	NUM		NUM	

【 図 1 7 】

パターン情報(パターン一致単語:標準形:変形候補単語)

対訳対ID	言語					
		ID	1	2	3	4
1	英	情報	(1-5 : 6 : 1-5)	(1-7 : 4 : 1-7)	(1-11 : 1 : 1-11)	(1-14 : 4 : 1-14)

【図 1 8】

	1-15	goals	goals	
	1-14	four	数詞	four
	1-13	with	前置詞	with
	1-12	half		half
	1-11	first	数詞	first
	1-10	the	冠詞	the
	1-9	in	前置詞	in
	1-8	advan- tage		advan- tage
	1-7	4	数詞	
	1-6	-		
	1-5	6	数詞	
	1-4	a	冠詞	a
	1-3	to	前置詞	to
	1-2	cruised		cruised
	1-1	Japan	固有名詞	japan
単語ID (1-単語)	単語	解析情報	読み	変数番号
言語	英			
文脈ID	1			

【図 1 9】

対訳辞書(日英)		
品詞	単語(日)	単語(英)
固有名詞	日本	Japan
名詞	前半	first half

【図 2 0】

対応付け情報				
対訳対ID	IID	単語ID(日)	単語ID(英)	対応付け手法
1	1	1-1	1-1	辞書
	2	2-1	1-11,1-12	辞書

【図 2 1】

対応付け情報				
対訳対ID	IID	単語ID(日)	単語ID(英)	対応付け手法
1	1	1-1	1-1	辞書
	2	2-1	1-11,1-12	辞書
	3	2-2	1-7	文字
	4	4-1	1-5	文字
	5	4-2	1-6	文字
	6	4-3	1-7	文字

【図 2 3】

最適対応情報	
対訳対ID	対応付け情報のIID
1	1,2,5,8,9,10

【図 2 2】

対応付け情報				
対訳対ID	IID	単語ID(日)	単語ID(英)	対応付け手法
1	1	1-1	1-1	辞書
	2	2-1	1-11,1-12	辞書
	3	2-2	1-7	文字
	4	4-1	1-5	文字
	5	4-2	1-6	文字
	6	4-3	1-7	文字
	7	2-2	1-7	パターン(p1,p3)
	8	2-2	1-14	パターン(p1,p3)
	9	4-1	1-5	パターン(p1,p3)
	10	4-3	1-7	パターン(p1,p3)
	11	4-3	1-14	パターン(p1,p3)

【図 2 4】

対訳対ID		
1	日本語	(((\$j1s 固有名詞)("は"助詞)("前半"名詞)(\$j2s 数詞)("点"接尾)("を"助詞)("奪い"動詞)(\$j3s 数詞)("-"記号)(\$j4s 数詞))
	英語	(\$e1s"cruised to a" \$e2s"-" \$e3s"advantage in the first half with" \$e4s"goals")
	変数箇所の対応	(((\$j1s \$e1s 固有名詞) (\$j2s \$e4s 数詞) (\$j3s \$e2s 数詞) (\$j4s \$e3s 数詞))

【図 2 5】

対訳対ID	言語	文
2	日	日本は最初4点を奪い6-4
	英	Japan cruised to a 6-4 advantage in the first half with four goals

【図 2 6】

言語	単語ID (文節・単語)	1-1	1-2	2-1	2-2	2-3	2-4	3-1	4-1	4-2	4-3							
		単語	は	最初	4	点	を	奪い	6	-	4							
英語	解析情報 読み	固有名词 nippon	助詞	名詞	数詞	接尾	動詞	動詞	数詞	記号	数詞							
		PN: 形																
英語	(1-単語)	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12	1-13	1-14	1-15		
		単語	cruised	to	a	6	.	4	advan- tage	in	the	first	half	with	four	goals		
英語	解析情報 読み	固有名词 japan		前置詞 to	冠詞 a	数詞		数詞		前置詞 in	冠詞 the	数詞 first		前置詞 with	数詞 four	goals		NUM: 数,パ
		PN: 語																
対応対ID		2																

【図 2 7】

対応付け情報				
対訳対ID	IID	単語ID(日)	単語ID(英)	対応付け手法
2	1	1-1	1-1	辞書
	2	2-2	1-7	文字
	3	4-1	1-5	文字
	4	4-2	1-6	文字
	5	4-3	1-7	文字
	6	2-2	1-7	パターン(p1,p3)
	7	2-2	1-14	パターン(p1,p3)
	8	4-1	1-5	パターン(p1,p3)
	9	4-3	1-7	パターン(p1,p3)
	10	4-3	1-14	パターン(p1,p3)

【図 2 8】

最適対応情報	
対応対ID	対応付け情報のIID
2	1,4,7,8,9

【図 2 9】

対応集		
対応対ID	言語	文
1	日	サッカー決勝ブラジルVSウルグアイ
	中	足球决赛巴西队对乌拉圭队

【図 3 0】

対応対ID	言語										
1	日	単語ID (文節-単語)	1-1	1-2	1-3	1-4	1-5				
		単語	サッカー	決勝	ブラジル	VS	ウルグアイ				
		解析情報	名詞	名詞	固有名詞	名詞	固有名詞				
		読み	sakka		buraziru		uruguai				
		変数候補			PN: 形,パ		PN: 形,パ				

【図 3 2】

パターン情報(パターン→単語:標準形:変数候補単語)

対応対ID	言語										
1	日	ID									
		情報									
	中	ID									
		情報									

【図 3 1】

パターン集

pID	言語	標準形	パターン	
p1	日	\$1	P単語ID	1
			単語	
			品詞	数詞
			変数候補	NUM

【図 3 3】

音韻辞書(中)

漢字	読み
足	zu
球	qiu
決	juè
賽	sai
巴	ba
西	xi
隊	duì
对	duì
乌	wu
拉	la
圭	gui
队	duì
...	...

【図 3 4】

対訳対ID	言語													
1	中	単語ID (1-単語)	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9	1-10	1-11	1-12
		単語	足	球	決	賽	巴	西	隊	対	烏	拉	圭	隊
		解析情報												
		読み	zu	qiu	jue	sai	ba	xi	dui	dui	wu	la	gui	dui
		変数候補												

【図 3 5】

固有名詞辞書(中)

巴西
乌拉圭
...

【図 3 6】

対訳対ID	言語													
1	中	単語ID (1-単語)	1-1	1-2	1-3	1-4	1-5	1-6	1-7	1-8	1-9			
		単語	足	球	決	賽	巴	西	隊	対	烏	拉	圭	隊
		解析情報					固有名詞			固有名詞				
		読み	zu	qiu	jue	sai	baxi	dui	dui	wulagui	dui			
		変数候補					PN: 辞			PN: 辞				

【図 3 9】

対応付け情報

対訳対ID	IID	単語ID (日)	単語ID (中)	対応付け手法
1	1	1-1	1-1,1-2	辞書
	2	1-2	1-3,1-4	辞書
	3	1-3	1-5	辞書
	4	1-5	1-8	辞書

【図 4 0】

対応付け情報

対訳対ID	IID	単語ID (日)	単語ID (中)	対応付け手法
1	1	1-1	1-1,1-2	辞書
	2	1-2	1-3,1-4	辞書
	3	1-3	1-5	辞書
	4	1-5	1-8	辞書
	5	1-3	1-8	音韻

【図 3 7】

パターン集

pID	言語	標準形	パターン	
p2	中	\$1	P単語ID	1
			単語	
			品詞	数詞
			変数候補	NUM

【図 3 8】

対訳辞書(日中)

品詞	単語(日)	単語(中)
名詞	サッカー	足球
名詞	決勝	決賽
固有名詞	ブラジル	巴西
固有名詞	ウルグアイ	乌拉圭

【図 4 1】

最適対応情報

対訳対ID	対応付け情報のIID
1	1,2,3,4

【図 4 2】

テンプレート

対訳対ID		
3	日本語	((“サッカー”名詞)(“決勝”名詞)(\$j1s固有名詞) (“VS”名詞)(\$j2s固有名詞))
	中国語	(“足球决赛”\$c1s “队对”\$c2s “队”)
	変数箇所の 対応	(((\$j1s\$c1s固有名詞) (\$j2s\$c2s固有名詞))

【図 4 3】

対応付け情報

対訳対ID	IID	単語ID (日)	単語ID (中)	対応付け手法
1	1	1-1	1-1,1-2	辞書
	2	1-2	1-3,1-4	辞書
	3	1-3	1-8	音韻

【図 4 4】

最適対応情報

対訳対ID	対応付け情報のIID
1	1,2,3